VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ

ÚSTAV RADIOELEKTRONIKY

FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION
DEPARTMENT OF RADIO ELECTRONICS

STRESS RECOGNITION FROM SPEECH SIGNAL

DOCTORAL THESIS

AUTHOR MIROSLAV STANĚK

SUPERVISOR Prof. MILAN SIGMUND

BRNO 2016

Abstract

Presented doctoral thesis is focused on development of algorithms for psychological stress detection in speech signal. The novelty of this thesis aims on two different analysis of the speech signal- the analysis of vowel polygons and the analysis of glottal pulses. By performed experiments, the doctoral thesis uncovers the possible usage of both fundamental analyses for psychological stress detection in speech. The analysis of glottal pulses in amplitude domain according to Top-To-Bottom criterion seems to be as the most effective with the combination of properly chosen classifier, which can be defined as language and phoneme independent way to stress recognition. All experiments were performed on developed Czech real stress database and some observations were also made on English database SUSAS. The variety of possibly effective ways of stress recognition in speech leads to approach very high recognition accuracy of their combination, or of their possible usage for detection of other speaker's state, which has to be further tested and verified by appropriate databases.

Keywords

Digital signal processing, speech signal processing, emotion recognition, psychological stress, formant, vowel polygons, glottal flow analysis, glottal pulse, Return-To-Opening phase ratio, Closing-To-Opening phase ratio, COG shift, classifiers, neural networks, Gaussian Mixture Models.

Abstrakt

Předložená disertační práce se zabývá vývojem algoritmů pro detekci stresu z řečového signálu. Inovativnost této práce se vyznačuje dvěma typy analýzy řečového signálu, a to za použití samohláskových polygonů a analýzy hlasivkových pulsů. Obě tyto základní analýzy mohou sloužit k detekci stresu v řečovém signálu, což bylo dokázáno sérií provedených experimentů. Nejlepších výsledků bylo dosaženo pomocí tzv. Closing-To-Opening phase ratio příznaku v Top-To-Bottom kritériu v kombinaci s vhodným klasifikátorem. Detekce stresu založená na této analýze může být definována jako jazykově i fonémově nezávislá, což bylo rovněž dokázáno získanými výsledky, které dosahují v některých případech až 95% úspěšnosti. Všechny experimenty byly provedeny na vytvořené české databázi obsahující reálný stres, a některé experimenty byly také provedeny pro anglickou stresovou databázi SUSAS.

Klíčová slova

Zpracování digitálního signálu, zpracování řečového signálu, rozpoznání emocí, psychologický stres, formanty, samohláskové polygony, analýza hlasivkových pulsů, *RTO* poměr, *CTO* poměr, COG posun, klasifikátory, neuronové sítě, Gaussovské smíšené modely.

Declaration

I declare that I have written my doctoral thesis "Stress Recognition from Speech Signal" independently, under the guidance of the dissertation supervisor and using technical literature and other sources of information which are all quoted in the dissertation and listed in the references at the end of the dissertation.

As the author of the dissertation, I, furthermore, declare that, as regards to the creation of this dissertation, I have not infringed any copyright. In particular, I have not unlawfully encroached on anyone's personal and/or ownership rights and I am fully aware of the consequences in the case of breaking Regulation §11 and the following Copyright Act No. 121/2000 Coll., and of the rights related to intellectual property right and changes in some Acts (Intellectual Property Act) and formulated in later regulations, inclusive of the possible consequences resulting from the provisions of Criminal Act No. 40/2009 Coll., Section 2, Head VI, Part 4.

Brno _____ _____

(author's signature)

Bibliographic citation

STANĚK, M. *Stress Recognition from Speech Signal*. Doctoral Thesis. Brno: Brno University of Technology, Faculty of Electrical Engineering and Communication, 2016. 122 pages.

ACKNOWLEDGEMENT

I would like to greatly thank to my supervisor Prof. Milan Sigmund for giving me opportunities and support in my research and for perfect supervising. My beloved Andrea, my family, my second family (Aleš, Petr, Jakub, Jiří & Jiří), my closest friends, my dog Max and my co-workers are also gratefully acknowledged for supporting me.

Brno, 31st May 2016

Miroslav Staněk

The research described in this treatise was performed in laboratories of the SIX Research Center, the registration number CZ.1.05/2.1.00/03.0072, the operational program Research and Development for Innovation.

List of Abbreviations

AANN	AutoAssociative Neural Networks
ADFJSS	Anger, Disgust, Fear, Joy, Sadness and Surprise
ANN	Artificial Neural Network
BTT	Bottom-To-Top
CDF	Cumulative Distribution Function
COG	Centre Of Gravity
DIF	Direct Inverse Filtering
DNN	Deep Neural Network
DT	Decision Tree
F	Formant
FFNN	Feed-Forward Neural Network
GerDA	Generalized Discriminant Analysis
GMM	Gaussian Mixture Models
GSS	Glottal Spectral Separation
GVV	Glottal Volume Velocity
GUI	Graphic User Interface
HMM	Hidden Markov Models
IAIF	Iterative and Adaptive Inverse Filtering
IFB	Bottom point of inflection
IFT	Top point of inflection
kNN	*k*-Nearest Neighbour
LF	Lijencrants-Fant
LFPC	Log-Frequency Power Coefficient
LLD	Low-Level Descriptor
LP	Linear Prediction
LPC	Linear Prediction Coefficient
LPCC	Linear Predictor Cepstral Coefficient
LSP	Linear Spectral Pair
MFCC	Mel Frequency Cepstral Coefficient
MLE	Maximum Likelihood Estimation

MLP	MultiLayer Perception
MSF	Modulation Spectral Features
PLPC	Perceptual Linear Prediction Coefficient
PNN	Probabilistic Neural Network
OSALPC	One-Sided Autocorrelation Linear Predictor Coefficient
OSALPCC	Cepstral-based OSALPC
SVM	Support Vector Machine
TEO	Teager Energy Operator
TTB	Top-To-Bottom

List of Symbols

α	Kurtosis
a_m	LPC coefficient
A	Frequency of higher formant
β	Skewness
β'	Decay constant
b	Current column vector feature
B	Frequency of lower formant
γ	Area
crl	Relative length of current glottal pulse COG
C	Controlling the exponential growth of sinusoid
CCV	Cross-correlation value
CTO	Closing-To-Opening phase ratio
δ	Dispersion coefficient
δg_{LF}	LF model
$\Delta\alpha$	Angle differences
$\Delta\alpha_{AVG}$	Average angle differences
Δd	Length differences
Δd_{AVG}	Average length differences
Δd_{min}	Minimal length differences
Δd_{max}	Maximal length differences
$\mathbf{d}$	Dispersion vector
d_M	Mahalanobis distance
dS_{AVG}	Average area differences
ε	Efficiency
E	Energy
E_0	Area balance scale factor
E_c	Efficiency coefficient
E_e	Value of glottal flow derivative
f	Frequency
f_s	Sampling frequency

F	F-ratio
F_i	Formant frequency
F_0	Fundamental frequency
G_p	Glottal pulse
θ	Angle of complex root pairs
H_0	Null hypothesis
H_1	Alternative hypothesis
i	Order number
j	Order number
k	Total selection number
L	Likelihood ratio
μ	Mean value of observed feature
μ	Reference column vector feature
M	Predictor order
n	Total selection number
n	Relative division
N	Total number of speakers
N_{cdn}	Number of correctly detected normal glottal pulses
N_{cds}	Number of correctly detected stressed glottal pulses
N_n	Total number of normal glottal pulses
N_s	Total number of stressed glottal pulses
p	p-value
prl	Relative length of current glottal pulse peak
$\boldsymbol{P}_{diff}$	Matrix of differences
R	Coefficient of variation
R_{AVG}	Average coefficient of variation
R_{plane}	Variation coefficient of actual formant plane
R_{shape}	Variation coefficient of actual vowel shape
RTO	Return-To-Opening phase ratio
σ	Standard deviation of observed feature
σ_{dS}	Standard deviation of area differences
S	Correlation matrix
t	Time
t_a	Time between glottal flow derivative minimum value and the end of its return phase

t_e	Time of glottal flow derivative
t_p	Time of glottal pulse maximum value
T	Fundamental period
T_c	Closing phase
T_{co}	Closed phase
T_{long}	Glottal pulse period for long spoken vowels
T_o	Opening phase
T_r	Return phase
T_{short}	Glottal pulse period for shortly spoken vowels
$\mathbf{v}$	Normalisation vector
w	COG shift
x	Set of features
$\bar{x}$	Average value of observed feature
X	Frequency value of important point
z	Complex roots

Contents

1. Introduction 1

2. Emotion analysis 2

2.1. Emotions 2

2.2. The State of the art 4

2.3. Emotion-oriented databases 8

2.4. Practical usage of automatic emotion recognition 11

2.5. Software applications 12

3. Stress 13

3.1. The State of the art 14

3.2. Stress-oriented databases 15

3.3. Practical usage and software applications 16

4. Technical Equipment 17

4.1. Microphones 17

4.2. Recording devices 18

4.3. Recording software 18

5. Doctoral Thesis Objectives 19

Used Speech Elements 20

6. Vowel polygons 21

6.1. Algorithms development 22

6.2. Research in speaker recognition 25

6.3. Psychological stress and vowel polygons 34

6.3.1. Stress division 35

6.3.2. Method application 36

6.3.3. Middle stress – experimental results 37

6.3.4. High stress – experimental results 39

6.3.5. Mixed stress – experimental results 41

6.3.6. Efficiency of vowel polygons 44

6.3.7. Summarization 46

6.4. Closure of vowel polygons 49

7. Formant changes varying on emotions 50

8. Glottal pulses 53

8.1. Mining the glottal pulses 54

8.2. Automatic estimation of glottal pulses and further filtration 60

8.3. Psychological stress detection 62

8.3.1. Return-To-Opening phase ratio in time domain 62
8.3.2. Experimental results 64
8.3.3. Closure of using *RTO* in glottis for stress detection 69
8.3.4. Top-To-Bottom Closing-To-Opening phase ratio in amplitude domain 70
8.3.5. Experimental results 72
8.3.6. Discussion 89
8.3.7. Bottom-To-Top Closing-To-Opening phase ratio in amplitude domain 90
8.3.8. Discussion 103

8.4. Closure of using *CTO*s and *RTO*s 103

9. COG shift 104

9.1. Maximum Likelihood Estimation 107

9.2. Statistical Testing 108

9.3. Discussion 110

10. Final Conclusion 111

References 114

List of figures

Fig. 1 Total information embedded in speech. 1
Fig. 2 Rectangular form of the wheel of emotions. 3
Fig. 3 List of emotions illustrated in two dimensional coordinate map. 3
Fig. 4 Speech features division. 5
Fig. 5 Possible equipment placement in room for secret speech record. 21
Fig. 6 Examples of /u/ vowel LPC spectrum 22
Fig. 7 Found /u/ vowel segments in word „osum“ by basic and improved algorithm. 23
Fig. 8 Developed software system: real vowel AIO vowel triangle in formant plane F1-F3 and its properties. 24
Fig. 9 The illustration of normalised and real small vowel triangle in the F1-F2 plane. 26
Fig. 10 Polar plots of vectors v for big and small vowel triangle in F2-F3 plane. 27
Fig. 11 Cumulative values of mean F4 obtained from five speakers in sets of 10 to 10000 speech frames of vowel /e/ 28
Fig. 12 Achieved area differences for AIO vowel triangle. 29
Fig. 13 The illustration of possible vectors d created by centroids of EIU12 vowel triangle. 30
Fig. 14 The distribution of COGs positions for EIO13 and EIO25 vowel triangles. 32
Fig. 15 The distribution of COGs positions for AEO12 and EIO34 vowel triangles. 33
Fig. 16 Differences of CGGs positions for normal and stressed speech in AIU23 vowel triangle. 34
Fig. 17 Differences between AEI34 created vectors' length and direction for middle, high and mixed stress level of 18 different speakers. 36
Fig. 18 Cross-correlation values over all vowel shapes for all used methods applied on middle stressed speech. 37
Fig. 19 Reached coefficient of variation in formant planes criterion applied on middle stressed speech 38
Fig. 20 Reached coefficient of variation in vowel shapes plane criterion applied on middle stressed speech 38
Fig. 21 Plane figuring out reached *R* for middle stress influence. 39
Fig. 22 Cross-correlation values over all vowel shapes for all used methods applied on high stress level. 39
Fig. 23 High stress level- reached coefficient of variation in formant planes criterion 40
Fig. 24 High stress level- reached coefficient of variation in vowel shapes criterion 40
Fig. 25 Plane figuring out reached *R* for high stress influence 41
Fig. 26 Cross-correlation values over all vowel shapes for all used methods applied on mixed stress influence. 42
Fig. 27 Reached *R* in formant planes criterion applied on mixed stressed speech 42
Fig. 28 Reached *R* in vowel shapes criterion applied on mixed stressed speech 43
Fig. 29 Plane figuring out reached *R* for mixed stress influence 43
Fig. 30 Experimentally obtained E_c values for middle stress influence by methods 1, 3 and 5 45
Fig. 31 Experimentally obtained E_c values for middle stress influence by methods 2, 4 and 6 45
Fig. 32 The derivations of /a/ vowel LPC spectrum and its important points. 50
Fig. 33 Relative positions of important points in F1-F2 and F2-F3 bands 51
Fig. 34 Illustration of glottal pulses series and its description. 54
Fig. 35 Flow charts of IAIF and DIF estimation algorithms. 55
Fig. 36 Glottal pulse differences depending on the speech tempo 56
Fig. 37 The shape varying of glottal pulses estimated by DIF and IAIF algorithms for different initialization number of possibly occured formants (from 1 to 3). 57
Fig. 38 The shape varying of glottal pulses estimated by DIF and IAIF for different initialization number of possibly occured formants (from 4 to 6) 58
Fig. 39 Shape differences of normalised glottal pulses estimated by DIF and IAIF at vowel's beginning and centre part 59
Fig. 40 The flow chart of glottal pulses automatic estimation and processing algorithm. 60
Fig. 41 Normalised glottal pulses before filtration estimated by IAIF algorithm. 61
Fig. 42 Normalised glottal pulses after filtration estimated by IAIF algorithm. 62

Fig. 43 Division of two-dimensionally normalised glottal pulse into *n*-percentage particular intervals of opening and return phase. 63
Fig. 44 Accuracy of stress detection depending on selected *n*-percentage interval and using *k*-Nearest Neighbour as a classifier. 65
Fig. 45 Accuracy of stress detection depending on selected *n*-percentage interval and using Random Forest Decision Tree as a classifier. 65
Fig. 46 Accuracy of stress detection depending on selected *n*-percentage interval and using Support Vector Machine as a classifier. 66
Fig. 47 Accuracy of stress detection depending on selected *n*-percentage interval and using Gaussian Mixture Models as a classifier. 67
Fig. 48 Accuracy of stress detection depending on selected *n*-percentage interval and using Probabilistic Neural Network as a classifier. 67
Fig. 49 Accuracy of stress detection depending on selected *n*-percentage interval and using Feed-Forward Neural Network as a classifier. 68
Fig. 50 The example of glottal pulse *n*-percentage division where selected closing phase is illustrated by dark grey and opening phase is represented by light grey colour. 71
Fig. 51 The example of differences varying on speaker‘s state in estimated glottal pulses by DIF in /u/ vowels’ beginning for ExamStress speaker 1 and 30% selected interval with average *CTO*s 72
Fig. 52 The flow chart of used psychological stress recognition algorithm. 73
Fig. 53 Reached efficiency for TTB *CTO*s, ExamStress database and GMM classifier. 76
Fig. 54 Reached efficiency for TTB *CTO*s, SUSAS database and GMM classifier. 76
Fig. 55 Reached efficiency for TTB *CTO*s, ExamStress database and FFNN classifier. 80
Fig. 56 Reached efficiency for TTB *CTO*s, SUSAS database and FFNN classifier. 81
Fig. 57 Reached efficiency for TTB *CTO*s, ExamStress database and PNN classifier. 82
Fig. 58 Reached efficiency for TTB *CTO*s, SUSAS database and PNN classifier. 83
Fig. 59 Reached efficiency for TTB *CTO*s, ExamStress database and SVM classifier. 84
Fig. 60 Reached efficiency for TTB *CTO*s, SUSAS database and SVM classifier. 85
Fig. 61 Reached efficiency for TTB *CTO*s, ExamStress database and DT classifier. 86
Fig. 62 Reached efficiency for TTB *CTO*s, SUSAS database and DT classifier. 87
Fig. 63 Reached efficiency for TTB *CTO*s, ExamStress database and kNN classifier. 88
Fig. 64 Reached efficiency for TTB *CTO*s, SUSAS database and kNN classifier. 89
Fig. 65 An example of *n*-percentage division for BTT criterion of *CTO* features. 90
Fig. 66 Reached efficiency for BTT *CTO*s, ExamStress database and GMM classifier. 91
Fig. 67 Reached efficiency for BTT *CTO*s, SUSAS database and GMM classifier. 92
Fig. 68 Reached efficiency for BTT *CTO*s, ExamStress database and FFNN classifier. 93
Fig. 69 Reached efficiency for BTT *CTO*s, SUSAS database and FFNN classifier. 94
Fig. 70 Reached efficiency for BTT *CTO*s, ExamStress database and PNN classifier. 95
Fig. 71 Reached efficiency for BTT *CTO*s, SUSAS database and PNN classifier. 96
Fig. 72 Reached efficiency for BTT *CTO*s, ExamStress database and SVM classifier. 97
Fig. 73 Reached efficiency for BTT *CTO*s, SUSAS database and SVM classifier. 98
Fig. 74 Reached efficiency for BTT *CTO*s, ExamStress database and DT classifier. 99
Fig. 75 Reached efficiency for BTT *CTO*s, SUSAS database and DT classifier. 100
Fig. 76 Reached efficiency for BTT *CTO*s, ExamStress database and kNN classifier. 101
Fig. 77 Reached efficiency for BTT *CTO*s, SUSAS database and kNN classifier. 102
Fig. 78 The illustration of the fundamental idea of COG shift feature. 105
Fig. 79 Differences of COG shift varying on speaker’s state for the first set of observed methods. 106
Fig. 80 Differences of COG shift varying on speaker’s state for the last set of observed methods. 107

List of tables

TABLE I Speech databases related to emotions 9
TABLE II Databases containing speech under stress 15
TABLE III Relative ratio of vowels occupied in speech 20
TABLE IV List of most suitable speech features for speaker recognition 25
TABLE V Experimentally achieved differences..... 26
TABLE VI Results of parameters testing used ANOVA 27
TABLE VII Average formant frequencies of Czech cardinal vowels used for the calculation of reference triangle areas 27
TABLE VIII Vowel polygons ranked by achieved coefficient of variation..... 29
TABLE IX Average coefficient of variation for top five vowel triangles and formant planes.... 29
TABLE X The list of vowel polygons ordered by minimal vector difference value..... 31
TABLE XI The list of vowel polygons ordered by dispersion coefficient..... 32
TABLE XII The best and the worst vowel polygons classified by each used method applied on middle stress..... 46
TABLE XIII High stress influence- the list of the top and bottom vowel polygons 47
TABLE XIV The best and the worst vowel polygons achieved by each used method for mixed stress level..... 48
TABLE XV Averaged real values of three return to opening-phase ratios in 5% selected interval, IAIF method, normalised sound vowel's beginning..... 63
TABLE XVI Final sorting of used types..... 68
TABLE XVII The efficiency of psychological stress detection reached by Method 1 and Method 2 74
TABLE XVIII Stress detection efficiency reached by Method 3 and Method 4 74
TABLE XIX Achieved psychological stress detection efficiency by Method 5 and Method 6 75
TABLE XX Stress recognition efficiency achieved by Method 7 and Method 8 75
TABLE XXI Average efficiency ε values over all 8 methods and both databases 77
TABLE XXII Average efficiency ε values over n percentage intervals in the range from 5 to 40 % 77
TABLE XXIII Average Stress Detection for ExamStress Database and FFNN Classifier..... 79
TABLE XXIV Average Stress Detection for SUSAS Database and FFNN Classifier 80
TABLE XXV Average Stress Detection for ExamStress Database and PNN Classifier..... 81
TABLE XXVI Average Stress Detection for SUSAS Database and PNN Classifier 82
TABLE XXVII Average Stress Detection for ExamStress Database and SVM Classifier 83
TABLE XXVIII Average Stress Detection for SUSAS Database and SVM Classifier 84
TABLE XXIX Average Stress Detection for ExamStress Database and DT Classifier 85
TABLE XXX Average Stress Detection for SUSAS Database and DT Classifier..... 86
TABLE XXXI Average Stress Detection for ExamStress Database and kNN Classifier 87
TABLE XXXII Average Stress Detection for SUSAS Database and kNN Classifier 88
TABLE XXXIII BTT *CTO*s - Average Stress Detection for ExamStress Database and GMM Classifier 91
TABLE XXXIV BTT *CTO*s - Average Stress Detection for SUSAS Database and GMM Classifier 92
TABLE XXXV BTT *CTO*s - Average Stress Detection for ExamStress Database and FFNN Classifier 93
TABLE XXXVI BTT *CTO*s - Average Stress Detection for SUSAS Database and FFNN Classifier.... 94
TABLE XXXVII BTT *CTO*s - Average Stress Detection for ExamStress Database and PNN Classifier 95
TABLE XXXVIII BTT *CTO*s - Average Stress Detection for SUSAS Database and PNN Classifier .. 96
TABLE XXXIX BTT *CTO*s - Average Stress Detection for ExamStress Database and SVM Classifier 97
TABLE XL BTT *CTO*s - Average Stress Detection for SUSAS Database and SVM Classifier 98

TABLE XLI BTT *CTO*s - Average Stress Detection for ExamStress Database and DT Classifier 99
TABLE XLII BTT *CTO*s - Average Stress Detection for SUSAS Database and DT Classifier... 100
TABLE XLIII BTT *CTO*s - Average Stress Detection for ExamStress Database and kNN Classifier 101
TABLE XLIV BTT *CTO*s - Average Stress Detection for SUSAS Database and kNN Classifier 102
TABLE XLV Experimentally Achieved Statistical Values of COG Shift 105
TABLE XLVI Statistical Values of COG Shift by MLE 107
TABLE XLVII Results of Statistical Testing 109

1. Introduction

Language is the unique communication instrument of plants, animals, humans and other beings and can be carried out by two basic ways- lingual and non-lingual. Speech can be defined as the lingual way of communication within humans and provided by vocal tract. Due to these facts, it is called as the vocalized form of human.

The cornerstone of word is phoneme which can be further divided into vowel and consonant group. Each word is created by the relevant set of phoneme units, exactly of their phonetic sounds. By the combination of spoken words, the speech is generated. The main reason of speech providing and generating is to transfer some information to audience by the vocalized form of language, but also another accompanying facts are transmitted. Figure 1 shows, what accompanying information is transferred. Content of spoken message is created in 75% by the object of conversation, exactly of the transmitted idea, and the least 25% of spoken message are referred to the speaker (15% to speaker identity and 10% to actual emotional state of speaker).

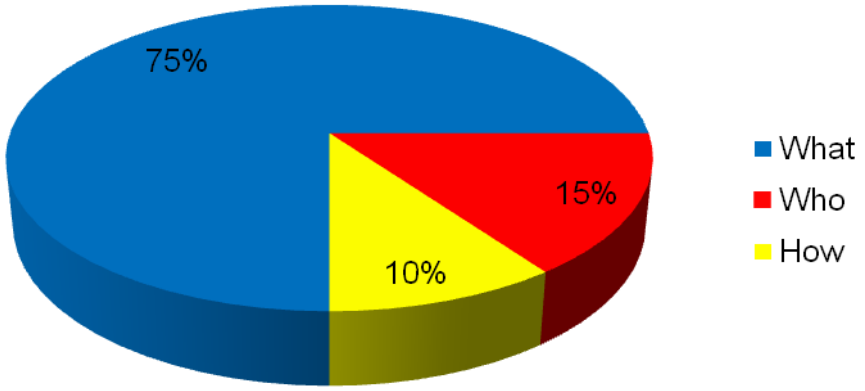

Fig. 1 Total information embedded in speech.

Nowadays, the speech signal is observed and processed in many areas:

- Speech recognition (voice recognition) is aimed on the converted linguistic content of speech to computer-readable format.
- Speaker recognition, further divided into speaker identification and verification, oriented on speaker identity recognition from speech signal used not only in forensic and security applications.
- Voice analysis which can be used in medical area (analysis of vocal tract to classify relevant dysfunctions) as well as in other application where the knowledge of the actual emotional state of speaker or alcohol/drug intoxication is necessary, e.g. call centres, intelligent house and nursing home, driving under drug/alcohol influence car protection etc.

- Speech synthesis is used for voice generation by computer or other similar equipment. This area can be used for interpretation of written text to spoken words to blind people using computer, navigation or machine reaction of given commands.
- Speech coding and speech compression are important areas in telecommunications to increase the amount of stored, transferred and heard information.
- Speech enhancement is the area of speech signal processing dealing with the improvement of the transferred speech signal quality by e.g. developing robust methods for noise reduction etc.

By previous list of speech signal processing usage, the information included in speech signal and possibly obtained is: speech and speaker identity as well as the actual emotional speaker mood, alcohol or drug intoxication and vocal tract dysfunction.

As the title of this thesis suggests, the main aim is the voice analysis oriented on determining the actual emotional state of speaker.

2. Emotion analysis

2.1. Emotions

Mentally or sociologically constructed processes containing subjective experiences of pleasure or displeasure can be defined as emotions [1] also influenced by hormones such as dopamine, serotonin, oxytocin, cortisol, and so on. Emotions are often disposed by the motivation, positive or negative, and followed by physiological changes, e.g. changes in heartbeat rhythm as well as in breathing and gesticulation etc., not only in human body. In the end of the 19th century, the importance of emotions in communication is mentioned in references [2] where Darwin argued that emotions are evolved by natural selection and pointed on the emotion occurrence in the animal world.

The term emotion is taken from French "émouvoir" which means "to stir up". For hundreds of years, many publications oriented on emotions have been written, mostly on psychological, sociological and other behavioural emotion roots. Basically, emotions are classified as reactions on internal or external events. Physiological, behavioural, neural and verbal mechanisms are included in these responses [3]. By Fox, emotion can be differentiated into:

- Feelings defined as subjective representation of emotions dependent on individual experience.
- Moods representing the long term duration of affective states and are less intense than emotions.
- Affect is used for description of emotions, feelings and moods together.

Emotions as discrete, measurable and psychological units are described in references [4] where six basic emotions have been classified. By Ekman, the basic emotions are: anger,

disgust, fear, happiness, sadness and surprise. The wheel of emotions have been further developed by Plutchik [5] for better illustration of opposite emotions and more punctual description and classification of six basic emotions laid by Ekman. New set of basic emotions (excited, happy, sad, angry, scared and tender) forms the full list of human emotion experience and the division into next eight kinds of relevant emotion is content for each basic emotion class. Rectangular form of the wheel of emotions is illustrated in Fig. 2.

Excited	Happy	Sad	Angry	Scared	Tender
Estatic	Fulfilled	Down	Irritated	Tense	Intimate
Energetic	Contented	Blue	Resentful	Nervous	Loving
Aroused	Glad	Mopey	Miffed	Anxious	Warm-Hearted
Bouncy	Complete	Grieved	Upset	Jittery	Symphatetic
Nervous	Satisfied	Dejected	Mad	Frightened	Touched
Perky	Optimistic	Depressed	Furious	Panic-Stricken	Kind
Antsy	Pleased	Heartbroken	Raging	Terrified	Soft

Fig. 2 Rectangular form of the wheel of emotions**.**

Another emotion analysis and classification has been defined by Schacter [6] where also the so-called emotion distances between individual emotional experiences have been mentioned. Emotion classes can be figured out in two-dimensional plane for better understanding of current feelings (see Fig. 3).

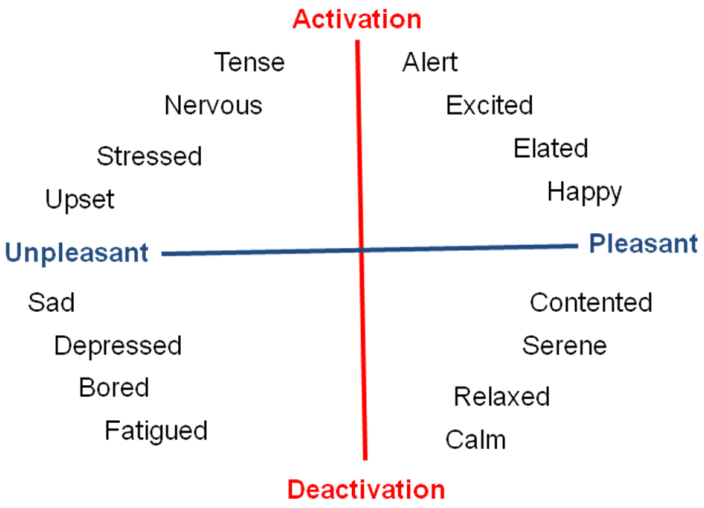

Fig. 3 List of emotions illustrated in two dimensional coordinate map**.**

For further needs of this thesis, classification into six basic emotions (anger, disgust, fear happiness, sadness and surprise) supplemented with stress, alcohol and similar toxic influences will be satisfied.

2.2. The State of the art

Actual methods used for emotion recognition from speech signal and papers published in recent years are described and listed in following section.

The general review of recognizing emotion in speech is presented in references [7] where the overview of used methods, speech features and obtained results is mentioned as well as the list of observed databases. In the case of emotion recognition, crucial decision is to choose the most suitable speech feature carrying unique information for each different state of speaker. Linear Prediction Coefficients (LPCs) and derivatives from LP residual can be mentioned as the basic and one of the most popular features for speech processing. The correlation of LP residual and the Glottal Volume Velocity (GVV) signals has been observed [8] which means the valid information of vocal tract producing speech are transmitted in LP residual signal. The higher order correlations of LP residual signal may be captured to some extent by using some helpful features, e.g. characteristics of GVV waveform or open and close glottis' phases, shapes of glottal pulse etc. [9].

Mel Frequency Cepstral Coefficients (MFCCs), Perceptual Linear Prediction Coefficients (PLPCs) and formant feature are the most used and mentioned speech parameters used for actual state of speaker recognition and so on [10].

Another possibly observed parameter of speech is so called prosodic feature which can be defined as speech features associated with larger units (syllables, words, phrases and sentences) and is often considered as suprasegmental information [7]. Acoustically, the prosody is specified by the duration patterns, intonation (base tone F_0) and energy. In references [7], the combination of used features for emotion recognition is listed as well as mainly used classifiers divided into two categories- linear (e.g. Naive Bayers classifier, Fischer's linear discriminant analysis, least square method, linear support vector machine) and non-linear, such as Hidden Markov Models (HMM), Gaussian Mixture Models (GMM), soft support vector machines, neural networks, decision trees and so on, classifiers.

One more written survey on speech recognition is mentioned in references [11] containing the list of used features, classification schemes, databases etc. The speech signal processing is also described in this paper in step-by-step for obtaining the best speech feature and the best results of desired application. Generally, the speech features can be divided into four main categories: continuous, qualitative, spectral, and TEO (Teager Energy Operator)-based features (see Fig. 4)

Continuous speech features are heavily used in determination of actual emotional state of speaker due to the most researchers believe much emotional content is carried by prosody continuous features, e.g. pitch and energy. This observation has been proved by several studies because the arousal state of speaker (high activation versus low activation) affects the overall energy, energy distribution across the frequency spectrum and the frequency and duration of pauses of speech signal [11].

The usage of some qualitative features of speech can be applied on emotion recognition. In references [11], some studies based on qualitative factors of speech, e.g. voice level, voice pitch, phrasing, temporal structures, breathing and so on, can be found. These studies are based on the approaching of speaker's vocal tract. It is necessary to mention the vocal tract can be better approached only in the case of glottal signal observation.

The third group of observed features is defined by the short time spectral features of speech because the impact of emotional state has been proved on the spectral energy distribution across the speech frequency range. In survey [11], the most used spectral features are LPCs, MFCCs, Linear Predictor Cepstral Coefficients (LPCCs), One-Sided Autocorrelation Linear Predictor Coefficients (OSALPCs), Cepstral-Based OSALPCs (OSALPCCs) and so called log-frequency power coefficients. The suitability of spectral feature for emotion recognition has been compared [12] for twelve speakers (six Mandarin and six Burmese) and six emotions (anger, disgust, fear, joy, sadness, surprise) and classified by HMM. The best average emotion recognition accuracy (77.1%) has been achieved by LFPCs followed by MFCCs (59.0%) and LPCCs (56.1%).

TEO-based features were introduced by Teager [13] and Kaiser [14] to prove the hearing is the process of energy detection. The nonlinearity of air flow in the vocal tract producing speech was observed experimentally [15]. Basically, the TEO-based features can be helpful in the case of emotion recognition because interactions between frequency components, fundamental frequency and harmonics, are included in TEO signal. This kind of feature, done experiments and achieved results are described more detail in references [11].

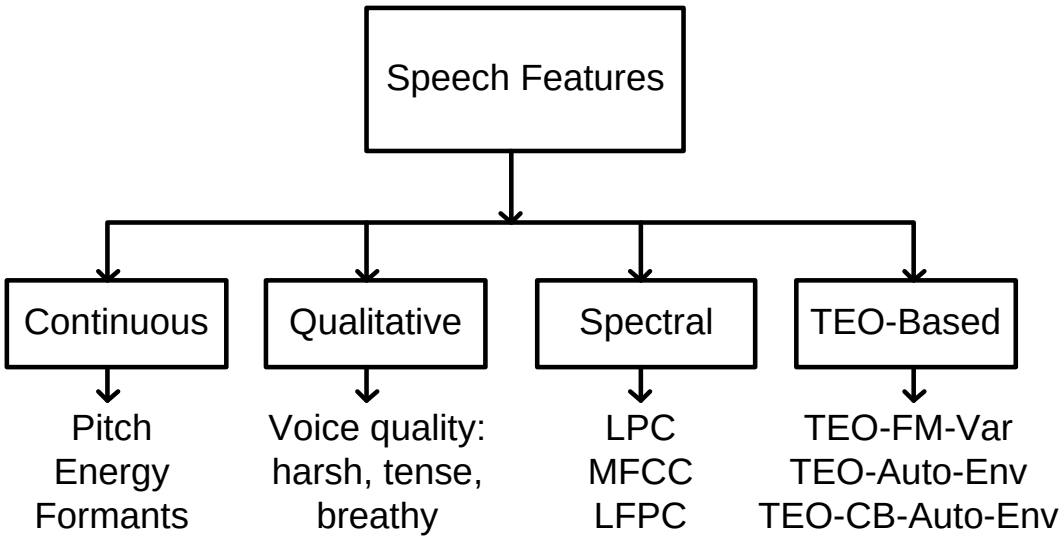

Fig. 4 Speech features division**.**

Introduction to the emotion recognition topic is also described in [16] where the basic of speech features, classification methods and previously published thematic works are written.

Following part of this section contains a brief survey of used speech features, classifiers, databases and achieved results in previous years.

In 2003, speech emotion recognition using HMM has been published by Nwe et al. [12]. As it can be obvious from the title of article, emotion states divided into six archetypes (Anger, Disgust, Fear, Joy, Sadness and Surprise- further ADFJSS) was classified by HMM. The LFPSs were chosen as the main speech feature and their accuracy reached the best results followed by MFCCs and LPCCs (see previous page). This experiment was applied on created speaker database containing 12 speakers and all emotional states were simulated. Another same titled paper [17] was written in 2001. In this paper, four different speech low level features (instantaneous energy, syllabic contour of energy, instantaneous pitch and syllabic contour of pitch) were used for emotion recognition separately and classified by HMM. The recognition method was applied on four different languages containing individually one female and one male speaker. All of recorded ADFJSS states filled by neutral mood were simulated and the results of automatic emotion classification were satisfied and similar to results reached by human judges.

Seven (ADJFSS filled with neutral speech) emotion states were also classified by low level speech features using the HMM and GMM [18]. The average accuracy 79.8% was obtained for acted emotions of five speakers performed in English and German. In another publication [19], the Hybrid GMM and Feed Forward Back Propagation Neural Network were used for emotion classification and HMM were used for speaker recognition. The MFCCs were used as the main feature in developed text dependent system. The experimental method was applied on created speaker database containing 25 speakers recorded in five different emotional states (happiness, anger, sadness, surprise and neutral mood). Achieved results in speaker and emotion recognition are more a less similar. The average accuracy is approached 90% for both recognition types. The best accuracy (approx. 96%) has been achieved for anger and normal speech in emotion and speaker recognition.

Four different types of statistical classification and the MFCCs set as the main speech feature were used in emotion recognition topic by Ayadi et al. [20]. Experimental study was applied on Berlin emotional speech database [21] containing six different emotions: anger, boredom, fear, happiness, sadness and neutral. The highest accuracy was achieved by GMM (76%), followed by HMM (71%), *k*-Nearest Neighbours (67%) and feed-forward neural network (55%). The efficiency of three different hierarchical classifiers was compared by Albornoz et al. and is described in [22]. By the variety of used speech features and classifiers for emotion recognition, interesting results were observed. The HMM are more suitable for classification of boredom, neutral and sadness, while the MLP (MultiLayer Perceptron) is better for discriminating between happiness, fear and anger. However, emotion influence can be embedded in some specific speech features, i.e. MFCCs plus velocity and acceleration coefficients well distinguish 3 emotional groups, whereas FV46 (12 Mean MFCCs, 30 Mean Log-Spectrum, $\mu(F_0)$, $\mu(E)$, $\sigma(F_0)$, σ (E)) recognizes better 2 emotional groups. Recognition method was also applied on Berlin emotional speech database consists of seven different emotion states.

Modulation Spectral Features (MSFs) were used for determining the actual emotional state of speaker in written paper [23]. The MSFs are divided into two groups: short-term and long-term MSFs, both describing the spectral behaviour. Furthermore, created system was applied on Berlin emotional speech database, received accuracy approximately 91.6%, and used the SVM as the classification tool. The Support Vector Machine and Berlin emotional database were also used for classification in [24] where different speech features were extracted from available utterances and five different emotional states (disgust, boredom, sadness, neutral and happiness) were used. The accuracy of developed system was approx. 66% using energy as

well as pitch, almost 71% using MFCCs and LPCs separately, and 82.5% using MFCCs and LPCs together.

The suitability of several different classifiers and speech features were observed by Schuller et al. [25] where emotional recognizing method was applied on created database (ADJFSS + neutral speech) containing 12 male and 1 female speaker speaking in German and English. The least error ratio (8%) was reached by a significant gain and the SVM was seemed as the most robust classifier for emotion recognition.

On German language, another developed automatic speaker-independent speech recognition system using HMM as classifier was applied by Vlasenko et al. [26]. The system recognition accuracy was verified by suitable combination of training and testing sequences which are performed by three different databases: the Kiel Corpus of Read Speech [27], VAM [28] and EMO-DB [29], and further used on specially created speaker database containing spontaneous emotions of 47 speakers, exactly TV talk show guests. By the introduced automatic speech recognition system, the cross-corpora classification performance reached approximately 72%. Another publication oriented on cross-corpora emotion recognition and unsupervised learning is written by Zhang et al. [30], where the usage of six different emotional speech databases is described. By using total 39 Low-Level Descriptors (LLDs) and SVM as a classifier, the suitability of unsupervised learning on cross-corpus emotion recognition has been evaluated. The similar set of six emotional speech databases were used for cross-corpora learning and emotion recognition by Jung et al. [31] where the developed automatic speech recognition system using classification by deep neural networks is described in more detail.

Deep Neural Networks (DNNs) are specified by more than one hidden layer and millions of free parameters. DNNs were used for standardization of Generalized Discriminant Analysis (GerDA) to learn discriminative low-dimensional features to fast emotion classification [32]. Better accuracy of proved experiments was reached by GerDA than previously used SVM by applying of developed system on nine available speech databases. In the case of emotion recognition, the application of fuzzy neural network using ARTMAP architecture is described in [33]. The set of 25 speech features including formants, pitch, energy, two first MFCCs and so on, is tested on created speaker database containing 30 speakers in neutral, happy and angry mood. Experimentally, the basic method reached accuracy of 84.97%. By the application of genetic algorithm for determining the optimal parameter value of used fuzzy network and implementation of fast correlation-based filter, presented system reached the average emotion recognition accuracy over 87.5%.

The achievement of formant feature frequency as the result of weighting the MFCCs and its further usage for emotion recognition can be found in [34].

To determine the actual emotional state of speaker, other types of classifiers are used. In [35], the emotion recognition efficiency is compared for different types of classification models related to the chosen set of speech features (e.g. zero crossing rate, energy, pitch, spectrum centroid etc.). Database containing seven speakers in six (ADJFSS) different emotional states was used for final comparison of recognizing level. It is argued the usage of Fischer criterion and SVM is the most suitable classification property for the robust automatic recognition system opposite of the usage of Principle Usage Analysis with Artificial Neural Network. Obviously, the best recognition results are obtained for anger and to distinguish between happiness and surprise reached the highest error ratio.

The application of binary decision tree as a classifier is explained in [36] where twelve MFCCs and their delta coefficients were used as the main speech feature for emotion recognition. Created system was applied on database containing four different states of speaker (anger, happiness, sadness and neutral) and based on existing databases: AIBO [37] and USC IEMOCAP [38]. The accuracy obtained by binary decision tree reached better results than using the SVM.

Emotion recognition based on decision tree using three different types of classifiers is described in [39] where the accuracy of created system is performed on German-spoken Emo-DB database and reached approximately 84% in distinguishing six different states of speakers. Techniques for spoken emotion recognition in call centres are described in [40].

As it has been mentioned, the prosodic feature can be also used for determining the emotional state. The combination of acoustic-prosodic information and semantic labels was utilized for recognition four different speaker moods (neutrality, happiness, anger and sadness) and approached efficiency of 83.55% [41]. Furthermore, segmental level prosodic analysis was used for classification of eight different emotions (ADJFSS + neutrality and sarcasm) and reached poor recognition results for usage word level prosodic feature (average 33%) as well as for usage prosodic feature of beginning (average 13%)/ mid (average 20%)/ end (average 26%) words [42].

Another usage of global and local prosodic feature as the main emotion recognition criterion is described in [43].

Further useful results and applied processes can be found in literature [44], [45], [46], [47], [48] as well as a survey on used spectral features [49], [50], [51] or possibly applied normalisation [52].

2.3. Emotion-oriented databases

Nevertheless some ordinary used emotional databases have been briefly outlined in previous subsection, the available and also ordinary non-available databases are listed in following rows. Databases can be simply divided into two main categories- containing act and real emotional states while the real states databases are mostly the product of long term data collecting.

Further, databases can be divided into multimodal, speech or face type.

Nowadays, the technical quality of speech records is more a less same for each individual database, so the biggest differences between databases can be seen in the number of recorded emotional states as same as the total number of speakers, their properties e.g. sex, age and spoken language- if native or not.

One note has to be mentioned. In the case of emotion recognition, archetypal and ordinary states of speaker should be contained in created database for correct training and testing recognition system. The survey of the most used emotional databases is presented in Tab. 1.

TABLE I SPEECH DATABASES RELATED TO EMOTIONS

Title of database	Created in	Emotions	Size	Language	Note
Van Bezooijen [53]	1984	anger, disgust, fear, joy, sadness, surprise, neutrality	4 phrases spoken by 4 male and 4 female speakers	Dutch	acted
Leeds-Reading [54]	1995-2001	range of full-blown emotions	4.5 hours	English	radio and TV interviews, real emotions
McGilloway, Cowie and Douglas-Cowie [55]	1996-1997	anger, fear, happiness, sadness, neutrality	5 passages spoken by 40 speakers	English	acted
Danish emotional speech [56]	1997	anger, happiness, sadness, surprise, neutrality	2 words, 9 sentences and 2 passages read by 4 speakers for whole emotional range	Danish	acted
Mozziconacci [57]	1998	anger, boredom, fear, disgust, guilt, happiness, haughtiness, indignation, joy, rage, sadness, worry, neutrality	three times repeated sentence by each actor in each mood	Dutch	acted
Abelin and Allwood [58]	2000	anger, disgust dominance, fear, joy, sadness, shyness, surprise	1 subject	Swedish	acted
Belfast structured [59]	2000	anger, fear, happiness, sadness, neutrality	20 passages spoken by 50 actors in each mood	English	extension of McGilloway database
Berlin [21]	2000	anger (hot), boredom, disgust, fear (panic), happiness, sadness (sorrow)	10 sentences spoken by 5 males and 5 females in each mood	German	acted

TABLE I (CONTINUED)

France et al. [60]	2000	depression, suicidal state, neutrality	48 females (10 therapists, 17 dysthymic, 21 major depressed) and 67 males (24 therapists, 21 major depressed, 22 high-risk suicidal)	English	real therapy sessions and call records
Iriondo et al. [61]	2000	desire, disgust, fury, fear, joy, surprise, sadness	a paragraph read in each mood by 8 speakers	Spanish	acted
Pereira [62]	2000	anger, happiness, sadness, neutrality	2 utterances read by 2 speakers in each mood	English	acted
CREST [63]	2002	wide emotional range	1,000 hours	English, Japanese, Chinese	the domestic and social interactions of volunteers through day
DARPA Communicator Corpus [64]	2002	frustration, annoyance	13,187 utterances (1,750 emotional)	English	simulated interactions with call centres
SYMPAFLY [65]	2003	joyful, neutral, emphatic, surprised, ironic, helpless, touchy, angry, panic	110 dialogues (29,200 words)	English	users booking flights via machine dialogue system
Yacoub et al. [66]	2003	neutral, anger, happiness, sadness, disgust, panic, anxiety, despair, elation, interest shame, boredom, pride, contempt	8 actors, total 2,433 utterances	English	acted
GEES [67]	2003	neutral, anger, happiness, fear, sadness	3 male and 3 female speakers	Serbian	acted

TABLE I (CONTINUED)

AIBO [37]	2004	joy, surprise, emphatic, helpless, touchy, anger, motherese, bored, reprimanding, neutral	51 children, total 5,822 words	German	child interaction with robot
Groningen ELRA corpus, no. S0020 [68]	2005	range of emotions	two short text read by 238 speakers in each mood	Dutch	acted
TALKAPILLAR [69]	2005	neutral, happiness, question, surprised, anger, fear, disgust, indignation, sadness, boredom	repeated 26 sentences by one speaker in three different activation levels for each mood	French	acted

The description of some databases listed in Tab. 1 and other previously used emotional speech corpora can be found in references [30], [31], [32].

2.4. Practical usage of automatic emotion recognition

The major reason for using the emotion recognition is to analyse the reaction of subject (person) on some object, resp. stimulus. In practise, the emotion recognition is mainly helpful in following branches:

Commercials - emotions are recognised for achieving the potential customer opinion and his feelings. This application of mainly visual recognition method (emotion recognition based on facial expression) is used daily by corporates, e.g. Google, Microsoft etc., for evaluation of commercial campaigns.

Public exposures, newspaper articles and social sites - recognising emotions in the public talk of popular persons (e.g. politicians), determining emotions from written text for achieving the unhidden opinion on laid situation, event, questionnaire etc.

Call centres - in the case of customer service or support, emotions are determined in call records for finding the solution of formed misunderstandings or mistakes to improve customer satisfaction. In rescue services as well as in **forensic applications**, emotion recognition is used for evaluation of the situation seriousness, for distinguishing between joke and real emergency call, for psychological analysis of caller and similar situations.

Health care – for establishing post traumatic disorders, depressions, suicidal thoughts, other psychological disorders, alcohol or drug intoxication, the emotion recognition can be used on delivered patients or accident participants.

Technical equipment – by embedding the suitable accessory into e.g. car, the driving under alcohol/toxic influence (recognized from speech) can be banned as well as the

reaction of steering wheel, throttle pedal etc. on negatively influenced (hot anger and similar moods) driver motions can be less sensitive and more under computer control for accident prevention. Another application of emotion recognizing can be found in embedding the whole algorithms into game console due the observation of the controversial computer game impact on child [70].

2.5. Software applications

As the actual topic of research, emotion recognition applications are mainly developed only in the form of classification algorithms, eventually tool kits, tested on accuracy and further improved.

Many developed software is described in publications. In 2000, the trio of developed applications was presented by Petrushin [71]. The presented set of developed applications is composed by **Emotion Recognition Game**, **Emotion Recognition software** for call centres and dialog emotion recognition program **SpeakSoftly**. Created tool kit called **Social Signal Interpretation** enabling the online emotion recognition is described in [72]. Another framework used for online emotion recognition from speech is called **EmoVoice** and is presented in [73] in more detail.

Of course, some commercial software exists. An application suitable for call centres is called **EMOSpeech** [74] and it records and analyses dialogues between costumers and operators, and finds the critical segments of dialogues where the customer was unsatisfied.

In the case of emotion classification, the **Praat** [75] software can be also helpful though it cannot be exactly mentioned as emotion recognizing application.

3. Stress

The word stress means tension, pressure and strain. Stress can be briefly defined as the state of organism during which the subject is faced to extraordinary conditions and can be divided into two types:

Eustress – the subject reaction on positive load stimulating the subject to better performance.

Distress – the overload which can cause disease, damage or, in the worst case, destroy the subject.

Obviously, stress is the psychic state of subject caused by external objects, so called stressors. Stressors can be divided into the five main groups of causing factors:

Physical – uncomfortable temperature, environmental noise and lightning.

Physic – responsibility, conscience, school or problems, frustration, age, unfulfilled expectations.

Social – negative personal relationships, unhealthy life style.

Traumatic – life events and situations (e.g. divorce, kidnapping, war, meeting etc.).

Children's – the reaction of subject on stress can be impacted by facing to stress in early age.

The influence of stress of subject's emotion can be observed in different ways. If the person is under depression, hormones produced by stress can be acted as sedatives leaded to deeper depression, low self-efficacy, guilt, self-hate, hopeless and helpless feelings. Depression states under stress are followed also by further problems such as: sleep problems, appetite changes, suicidal thinking etc.

States of mania or depression can be also powered as the result of long lasting stress in the case of bipolar disorder, where joy, anger, guilt and hopeless feelings are impacted at most. Of course, stress in bipolar disorder is followed by other symptoms like rapid speech (thoughts faster than spoken words), chronic tiredness and sleep problems.

Generally, it can be said that negative emotions are multiplied by the stress influence. By previous rows the relation between stress and emotions has been introduced and, for recapitulation, expressing the emotions states like anger, guilt and joy is mostly affected by present stress. In the majority of written publication, stress and emotions have been investigated separately though the relation between stress and emotions is obvious. Due to the theory of physical stress is tantamount as the theory of emotions, these two fields can be observed together.

3.1. The State of the art

Following subsection is presented as an equivalent of subsection 2.2., where the previously achieved results and used methods for stress recognition in speech signal is going to be listed. Same conditions are validated as in the case of emotion recognition, so this subsection presents only important notes.

A survey on stress recognition in speech signal in can be found in [76] where all previously used techniques are described.

Best results in classifying speech under stress were obtained by the nonlinear model of the phonation process filling by the spectral distribution of the glottal energy [77]. Lech and He applied their recognition methods on database containing 7 speakers under stress (3 female and 4 male speakers). On so called SUSAS corpus, recognition method used 39 speech parameters related to MFCCs as the main feature, HMM as a classifier and SVM for training and adaptation functions are described in [78], where the recognition efficiency of value approaches 95% is also mentioned. Hidden Markov Models were also used as a classifier by Hansen et al. [79]. In written paper, other methods of stress classification, difference in suitability between different speech features, previously published results and so on, are described by Hansen. Another type of spectral analysis and the study of empirical model both used for stress recognition are described in more detail in [80].

The advantages and disadvantages of using wavelets in stress detection are described in detail in [81] as well as emotion detected by using the wavelets. New harmony features based on music knowledge is introduced by Yang et al. [82] where the generation on base tone harmonics are described as a dependency on actual state of speaker, exactly on stress influence.

In references [83], the suitability of higher frequency parts of spoken vowel for stress detection is described. Other observations have been attempted for achieving the most suitable indicator of stress occupation in speech. Under the stress influence, the changes in fundamental frequency [84] were published as well as pitch changes [85], vowel duration and formants position [86]. The combination of Teager energy values and MFCCs [87] were used for stress classification as well as another combination containing Teager energy and slope of glottal spectra [88].

3.2. Stress-oriented databases

A survey on previously used databases containing recorded speech under stress is occupied in this subsection. Obviously, the list of stress oriented databases is going to be more poor in the comparison with databases containing emotions because of the stress cannot be acted and all records are originated from real situations. Nevertheless, the list of previously published speech under stress databases is shown in Tab. 2.

TABLE II DATABASES CONTAINING SPEECH UNDER STRESS

Title of database	Created in	Situation	Size	Language	Note
SUSC-0 (ground-to-air) [89]	-	Military communica-tion. Fighter pilots in ascent	11 males, record of length 15 minutes for each speaker	English	Non-native speakers
SUSC-0 (Aircraft crash) [89]	-	Ejecting off the aircraft	23 minutes	English	Poor quality
SUSC-0 (F-16 Engine out) [89]	-	Successful engine-out landing	15 minutes	English	-
SUSC-1 (Physical stress) [89]	-	Fair physical load (running up and down three floors)	10 males, 10 females, 10 times repeated 2 sentences in 10 different days	English	Phone quality.
Tolkmitt and Scherer [90]	1986	Answers on questions	33 males and 27 females	German	Three vocal responses
SUSAS [91]	1998	Various situations.	13 females, 19 males, 16,000 words	-	Real and simulated stress.
Rahurkar and Hansen [92]	2002	-	6 soldiers	English	Five stress levels
Scherer et al. [93]	2002	Normal and stress condition	100 speakers, 2 tasks spoken by each speaker	English and German	Effects on the stress and load
McMahon et al. [94]	2003	-	29 speakers	English	-
Fernandez and Picard [95]	2003	Responses on mathematica l problems	4 subjects	English	-
ATCOSIM [96]	2008	Speech of air traffic control operator	10 speakers, length 10 hours	English	Non-native
IEM-PSD [97]	2013	Communica-tion with pilots	> 700 utterances, > 7 hours	English and German	Various stress level and situations

3.3. Practical usage and software applications

Similar to emotion recognition systems, the market oriented on software detecting stress in speech signal is limited. In references, developed algorithms are published and are still in testing phase. Obviously, the usage of stress detection is similar to emotion recognition, as it has been mentioned earlier, these both phenomena have more a less same base and can be influenced by each other.

In the most publications e.g. [79], [98], [99], only algorithms have been developed in the case of providing research and not further presented in the form of independent software. Nevertheless some commercial applications recognizing speech under stress are created and available for practical usage. The set of software analysing speech was developed by company **Nemesysco** [100]. This set can recognize five different stress types and is composed by **LVA 6.5** (professional investigating tool), **RA7** (prevention and cheating detection), **LVA-i-CR** (judging of trust risks), **LVA-i-HR** (human resources control tool) and **QA5** (investigation of call-centre customer satisfaction) software. Software used for stress and emotion analysis is created by **AGI** company [101] and the software line is marked as **Sensibility Technology**. In the case of stress detection as well as emotion recognition, commercial products called **ST Emotion SDK** and **ST-CRM** can be used for speech analysis. Other commercial programmes created for speech analysis are **X13-VSA** [102] (available in Home, Pro and Cobra version) used for lie detection and voice stress analysis systems **AVSAPRO** [103].

Practical usage of stress detection can be following:

Lie detection – to observe and classify if the subject is cheating or not telling the truth.

Mental analysis – to observe if the subject can be possibly dangerous (psychopath, criminal etc.) and if the actual situation is controlled by subject (fighter pilot, personal driver and so on).

Health care – stress, exactly in the form of Post-traumatic stress disorder, can lead to heavy depressions, psychical problems and suicidal thinking. By these reasons, stress is analysed not only in the case of military health care [104].

Detail description and understanding of emotions and stress occupied in speech can be found in publications like An Argument for Basic Emotions [4], Handbook of Emotions [105], Three Dimensions of Emotions [106] and Emotion: A Psychoevolutionary Synthesis [107].

4. Technical Equipment

Sound capturing is an important part of speech signal processing. Low quality records can negatively influence analysing, processing and classifying. Moreover, badly captured sound records can be absolutely useless due to missing frequency part of records. This starting part of speech signal processing is provided by technical equipment briefly described in following subsections.

4.1. Microphones

In the case of emotion recognition and stress detection in speech and stress, many requirements exist on microphone for sound capturing in the most natural way. Main properties of used microphone can be defined as follows:

Omnidirectionality – the first requirement on possibly used microphone is a satisfactory sound capturing in the whole room or open area without any needed manipulation with microphone. Due to this reason, microphones described by omnidirectional, cardioid or possibly bidirectional polar pattern are suitable for speech signal recording. These patterns are typically used in condenser microphones characterized by almost flat frequency characteristic and perfect sensitivity on sound capturing of distant sound source. Condenser microphones are fed by source producing voltage of value 48V and are suitable for infrasonic applications but the characteristics in low frequencies are not listed by the manufacturers. Their microphone capsules are suitable for embedding into microphone fields to reach better directionality.

Infrasonic sensitivity – it is proved the useful information generated by glottis is transmitted in infrasonic frequency range (units of Hz). Owing to this fact, used microphone has to be sensitive in low sound frequencies and are mostly created by carbon/capacitor capsule with flat low frequency characteristic, optical parts or mechanically fixed accelerometers. In the comparison with conventional microphones, infrasonic microphone can be hidden in a longer distance away from source due to better low frequencies propagation and their less attenuation. Infrasonic microphones are used in the field of lie detection, measuring the influence of environment on health, seismic and volcano monitoring [108] and so on. As the suitable microphone for infrasonic recording, it can be mentioned precisely made condenser microphone G.R.A.S. 40AN disposing with cut-off frequency lower than 0.5 Hz [109] also used in speech processing experiments oriented emotion analysis [110]. The usage of piezo based microphones is also possible.

Wireless transmission – well hidden, possibly small, microphone without cord is necessary for capturing the natural behaviour of subject. Long lasting functionality on battery source is also desired.

By previous list of requirements on microphone, it is obvious the right choice of microphone might be difficult but all advantages can be achieved by using combination of all three microphone groups and multichannel recording interface together.

4.2. Recording devices

For conversion of captured analogue speech signal to digital form, suitable recording devices have to be used. Following subsection is oriented on the description of possibly used recording equipment divided into three main groups:

Portable devices – so-called all-in-one devices containing integrated simple sound card, software and mostly one microphone input. The length of records is limited by used memory card. Due to frequency characteristic of whole portable device, portable recording device is most suitable for capturing audibly part of speech signal. Its finishing is also suitable for secret record purchasing.

Recording stations – can be defined as a compromise between portable devices and computer powered stations. Higher number of independent inputs is typical for recording stations as well as limited recording software and storing records to memory cards.

Computer stations – good quality records can be captured by the combination of audio interface controlled by computer. Used audio interfaces can be divided into three groups: internal sound card, internal sound card with external module and external sound card. The connectivity with computer is realized by PCI bus for internal cards or by USB/FireWire for external sound cards. Advantages can be viewed in the perfect quality of sound recording, number of input channels with independent controlling, connectivity to various devices and the pairing with professional recording software. Owing to the musical usage of sound cards, the frequency range starts at 20 Hz but the usage in infrasonic range has been practically observed. In the group of professional recording interfaces, external sound card **Sound Devices USBPre2** has been applied on seismic observations [108] though its frequency range stated by manufacturer starts at 10 Hz.

4.3. Recording software

In the case of sound recording, computer stations are controlled by recording software. More input channels can be captured together independently by using the suitable combination of recording software, sound card and its drivers which can cause better records quality and improve recognition ability. The list of commonly use recording software can be content of **Cubase** and **Nuendo** (both produced by company Steinberg), **ProTools** and **Cakewalk Sonar**. The differences between these programmes can be seen in GUI and higher level functions. Obviously, the core of all recording software is defined uniformly and containing similar functions.

5. Doctoral Thesis Objectives

By previous sections, existing concepts in the field of emotion recognition were introduced. Nowadays not only in the case of speech signal processing, presented and known researches are further observed for achieving better accuracy, higher efficiency or for finding new more effective way leaded to the same aim. Written doctoral thesis is presented for outlining possible new methods recognizing emotions/stress in speech.

Hence, the doctoral thesis aims were laid as follows:

- To create suitable speech database.

  The first idea of this doctoral thesis was to create appropriate speech database of six different emotional states and alcohol intoxication for Czech female and male native speakers. But the real conditions are not suitable for creating this large database, because it is very difficult and time-consuming to capture real emotions for the highest number of speakers. Due to this reason, our experiments are only focused on real stressed and normal states of speakers which were recorded during appropriate situations at our department, because in general it is necessary to observe the impact of novel approaches on real not on acted emotions.

- To develop algorithms for obtaining desired speech features and analysing speech.

  As it was mentioned, the suitability of each speech feature is different for each purpose. By observations, highest speech feature differences depending on spoken emotion have to be found for stress recognition within the created speaker database. The glottal pulses have to be also observed because so much useful information of actual state of speaker is occupied in them. In this case, the main emphasis will be oriented just on the glottal pulse analysis. Simply, the combination of used speech features and glottal pulses behaviour is going to be observed.

- To develop methods for stress recognition.

  By setting suitable classifier on previously obtained speech features, emotions will be recognized. Obviously, both methods (analysing and recognizing) should be robust and speaker independent. In that case, the possibilities of speaker recognition have to be also observed and developed speaker recognizing algorithms will be further modified and applied on stress detection. The efficiency of created speech processing system will be compared with other available products.

Used Speech Elements

Presented research is mostly oriented on Czech language and vowels. Vowels are the special type of spoken phonemes characterized by the periodical signal form. Vowels are also generated by free air flow resonating in relevant cavities. Though the forty different spoken phonemes exist in Czech language, speech is consisted of vowels in 41.377 % (see Tab. 3) and almost every single word contains at least one vowel. Due to these reasons, research based on vowel properties can be performed.

TABLE III RELATIVE RATIO OF VOWELS OCCUPIED IN CZECH SPEECH [111]

Phoneme	Occupation [%]	Phoneme	Occupation [%]
/e/	9.216	**/é/**	1.182
/o/	7.904	**/ú/**	0.919
/a/	6.189	**/ou/**	0.659
/i/	6.164	**/au/**	0.030
/í/	4.571	**/eu/**	0.015
/u/	2.369	**/ó/**	0.011
/á/	2.148		

6. Vowel polygons

Fundamental observations were aimed mostly on the suitable speech feature extraction via created algorithms and on observation in the field of speaker recognition leading to investigation of most uniform vowel polygons within created speaker database representing the normal state of speaker at most. The basic observations were also done in emotion recognition. Provided research is described in following subsections.

Firstly, equipment used for recording speech signal is consisted of two-channel external USB sound card Line6 UX2 (substituted in some cases by eight-channel equivalent Line6 UX8), condenser microphone Omnitronic IC-1000 PRO (characterised by cardioidal polar pattern and is suitable for overhead sound recording of whole room), wireless microphone Shure SM58 (popular vocal dynamic microphone characterised by faithful speech reproducing and can be hidden due to wireless signal transmission), notebook Gigabyte U2442N and recording software Nuendo, version 4.3. Possible equipment placement in room for secret speech recording is illustrated in Fig. 5.

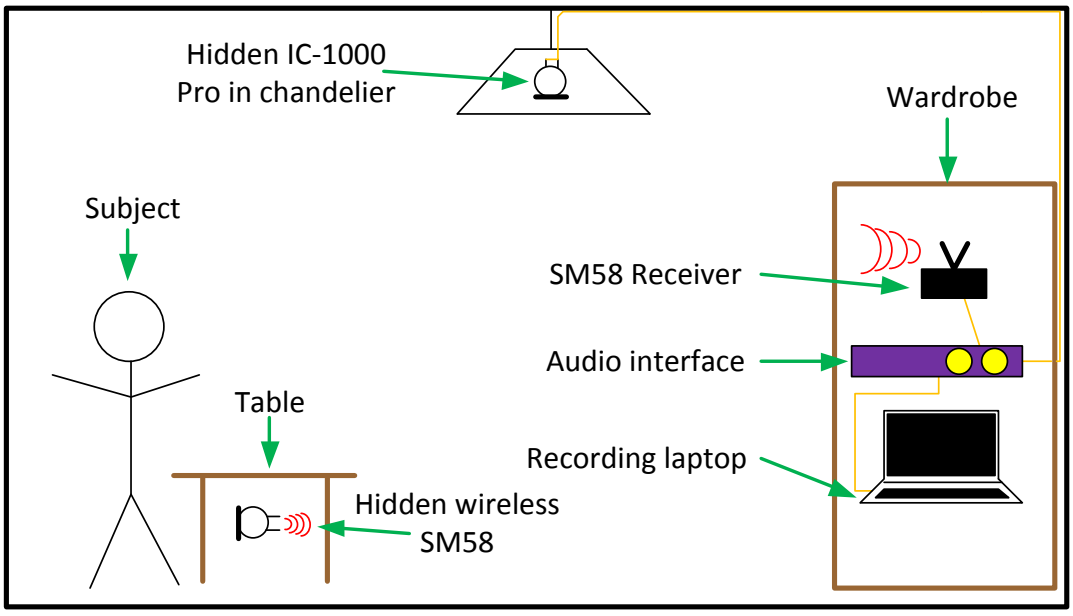

Fig. 5 Possible equipment placement in room for secret speech record.

6.1. Algorithms development

The first step into the emotion recognition topic is the development of vowel detecting algorithms in fluent speech because obtaining pure and true data is very important request for further processing.

In firstly provided experiments, the vowel recognition was based only on the position of first two formants determining the spoken phoneme. In Fig. 6, the examples of 17 different speakers /u/ vowel spectra are illustrated for showing the fact of spoken phoneme determining by the position of formants F1 and F2, where ordinal formant frequency intervals are performed by grey shadow area. Software was created only in the form of MATLAB based console application with also developed extension modules (mainly statistics modules). The created software system was presented at international student conference POSTER and its details and description can be found in relevant proceedings [112].

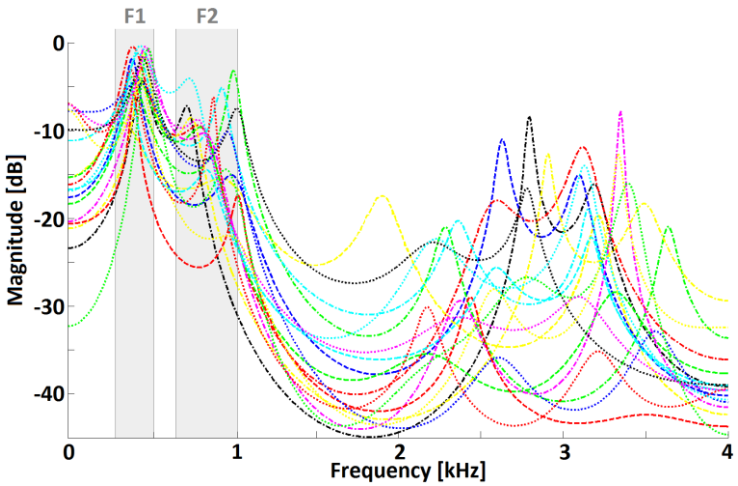

Fig. 6 Examples of /u/ vowel LPC spectrum (17 different speakers).

In the case of relatively high ratio of parasitic (false detected) vowel segments, the improvement of original algorithms was necessary. The peak error ratio was reached for /i/ vowel and it approached 40 % including false detected segments because of the relatively high leak between formants F1 and F2 for /i/ vowel possibly colliding with surrounding noise and parasitic environmental sounds. Due to this fact the original software had to be modified. By statistical observations of results received by first software version, almost all false detected vowel segments were located in the non-speech part of record and were suited in the small groups of max segments number 2. For this reason, the total length protection of found vowel segments had been implemented into original software.

Mentioned improvement is the form of retroactive checking of previously found vowel segments and it helps for erasing false detected segments and finding miss-detected vowel short-length section between two vowel parts. By this retroactive checking, the error ratio has been decreased by 38.8% at average.

The impact of improved algorithm (green boundaries) on found vowel segments by so-called basic algorithm (red boundaries) is illustrated in Fig. 7 where the signal form of Czech word “osm” (spoken as “osum”) is also figured out. Obtained results and cores of algorithms were presented at international conference on Telecommunication and signal processing [113].

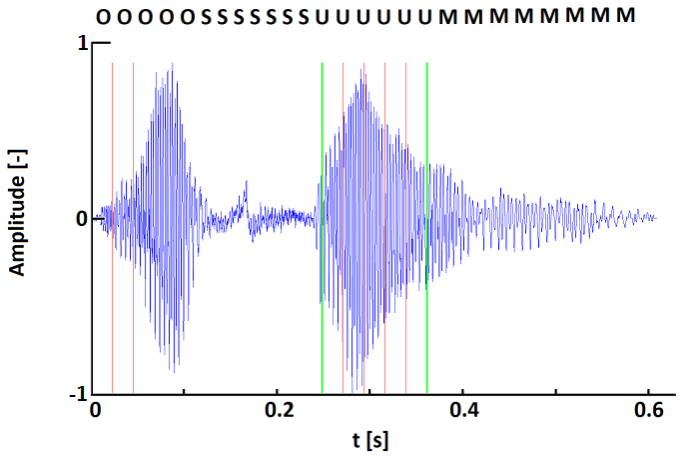

Fig. 7 Found /u/ vowel segments in word „osum“ by basic (red lines) and improved (green lines) algorithm.

Further, new software system for vowel recognition in fluent speech was created for usage in speaker and emotion recognition field. This application contains Graphic User Interface (GUI) and used two levels recognition tool for vowel detection. Thirteen MFCCs, 13 velocity (delta) and 13 acceleration (delta-delta) coefficients are used as a classification speech feature in both recognition levels. Reference values of observed features were mined from created speaker database containing 13 male and 12 female speakers.

The first level recognition tool is based on speech segment classification using Mahalanobis distance between reference and current pattern. Mahalanobis distance can be defined as follows

$$d_M(b) = \sqrt{(b-\mu)^T S^{-1}(b-\mu)}, \tag{1}$$

where b is column feature vector of current speech segment, μ is column vector of reference feature values and S stands for correlation matrix.

The second recognition approach is based on feed-forward neural network with eight hidden neurons. By suitable network training, the accuracy of vowel recognition was reached 97%. It is necessary to mention that, both recognition levels work and are retroactively checked independently. Finally, current vowel segment belongs into one of five vowel groups (/a/, /e/, /i/, /o/, /u/) if result detected by both recognition levels is the same. In the other case, current segment belongs into the sixth group containing only other phonemes. Formant frequencies are determined from found vowel segments directly from its LPC coefficients. This method is based on relation between formant frequencies and complex roots $z=re^{\pm j\theta}$ of the equation

$$1-\sum_{m}^{M} a_m z^{-m} = 0, \tag{2}$$

where a_m are LPC coefficients and M stands for predictor order. Then, the formant frequency F is defined as

$$F = \frac{f_s}{2\pi}\theta, \tag{3}$$

where f_s is sampling frequency of speech signal and θ is the angle (in radian) of the complex root pairs. The order of used linear prediction was set to M=10. More details about the algorithm can be found in [114].

The statistics tool is implemented in mentioned software system for showing parameters of found vowel formants and saving them into external data files for further possible processing. The main function of developed software system is to present obtained vowel data graphically in progressive view- so called vowel polygons. Vowel polygon is defined by desired vowels' apexes of coordinates set by reached average formant values. All possible vowel polygons (ten vowel triangles, four vowel tetragons and on pentagon) can be figured out in ten, according to total number of possibly existing formants in spectra, different formant planes. The formant plane can by defined as two dimensional space generated by firstly and secondly chosen formants values as horizontal and vertical axes. Real situation achieved by developed software system is illustrated in Fig. 8, where the AIO vowel triangle in formant plane F1-F3 is shown. As it can be seen, developed software system shows results and properties (symmetry, area, centre of gravity position etc.) of desired vowel polygon by relevant table.

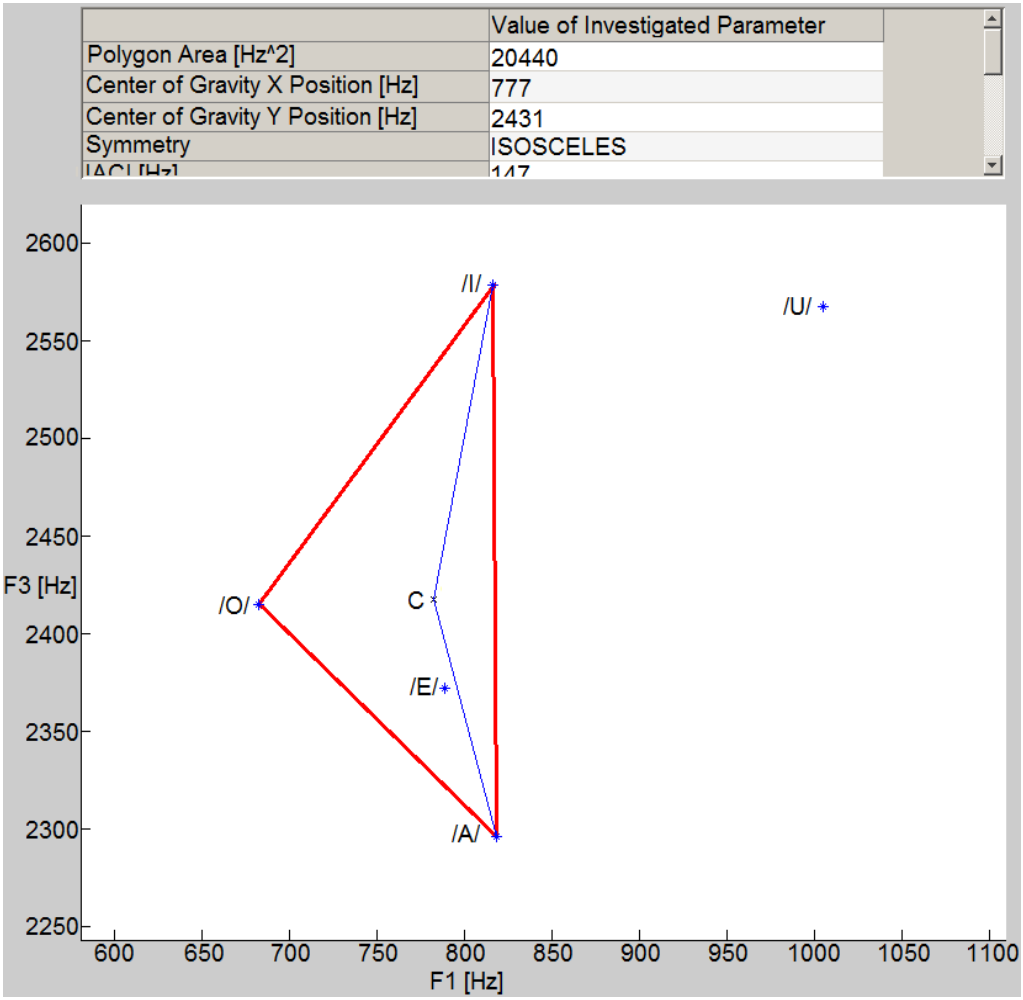

Fig. 8 Developed software system: real vowel AIO vowel triangle in formant plane F1-F3 and its properties.

The main idea of creating vowel polygons is based on possible graphical expressions of different emotional speaker states, speaker identities and so on.

Developed software system was presented at international conference on Telecommunication and Signal Processing 2014 and it is described in more details in relevant proceedings [115].

6.2. Research in speaker recognition

Following subsection brings observations received in the field of speaker recognition.

First experiments were oriented on finding the most suitable speech feature possibly used for speaker recognition. The set of observed speech features was consisted of: four formants, four formant bands, eleven LPCs, ten LPSs and thirteen MFCCs. For created speaker database containing 12 Czech native speakers reading same text, the set of observed speech features was mined by the created text-independent software described in [112]. The suitability of each speech feature was observed for each individual vowel separately and its uniformity/exclusivity within speaker database was classified by statistical methods including *F*-ratio simply defined as follows

$$F = \frac{Variance\ of\ speaker\ means}{Mean\ intraspeaker\ variance}, \tag{4}$$

leading to its advance definition by which the *F*-ratio of each feature can be calculated as

$$F = \frac{(n-1)\sum_{i=1}^{k}(\bar{x}_i - \bar{x})^2}{(k-1)\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x})^2}, \tag{5}$$

where x_{ij} is the feature value of *j*-th speaker during the segment *i*, *k* is the total selection number, *n* is the total number of current selection and $\bar{x}$ is an average value of observed feature. By previous equation leads, if the value of observed features will be more a less equal, result will reach the value near 1. But if the value of observed speech features will be different within the speakers, the *F*-ratio will achieve higher values dependant on the difference of speech feature within speaker database.

TABLE IV LIST OF MOST SUITABLE SPEECH FEATURES FOR SPEAKER RECOGNITION

Parameter	**/a/ [%]**	**/e/ [%]**	**/i/ [%]**	**/o/ [%]**	**/u/ [%]**	$\sum_{avg}$ **[%]**	**Rank**
LSP6	-	-	128	170	162	92.0	1.
LSP10	47	146	117	-20	-	72.5	2.
LSP8	-	-	86	-	16	51.0	3.
LSP7	-	-	-	-	46	46.0	4.
LSP9	-	45	-	-	-	45.0	5.
LPC8	7	41	-	-	-	24.0	6.
F4	-	9	-	15	-	12.0	7.
LSP5	-	-	-	69	-51	9.0	8.
LPC3	35	-	-46	-12	45	5.5	9.
LPC2	-	-	-89	-	96	3.5	10.
...							
MFCC6	-	-102	-	-	-	-20.4	23.

In the case of feature speaker variability, received values of *F*-ratio were observed by the set of tests, e.g. Bonferroni etc. The suitability of each speech feature for speaker recognition is listed in Tab. 4 in order. Obviously, the most suitable speech feature for speaker recognition

is the sixth Linear Spectral Pair (LSP6) followed by LSP10 and so on. Detailed description of provided tests, observations and achieved results can be found in [116]. But briefly, on the thousands of found and separated vowel segments, observed features were mined and statistically ranked for each vowel separately. Exactly the relative deviations of calculated *F* and *t* ratios were summed for each vowel and ranked from the highest to the lowest value.

Further these partial results, exactly the calculated standard deviations, are summed over the occurred vowels and then the mean value over all possible vowels $\sum_{avg}$ is calculated. These final results are further ranked by the mean standard deviation value from the highest to lowest value, because the highest values as possible are necessary to express to lowest speaker uniformity depending on actual feature within the database, which is wanted in the case of speaker recognition.

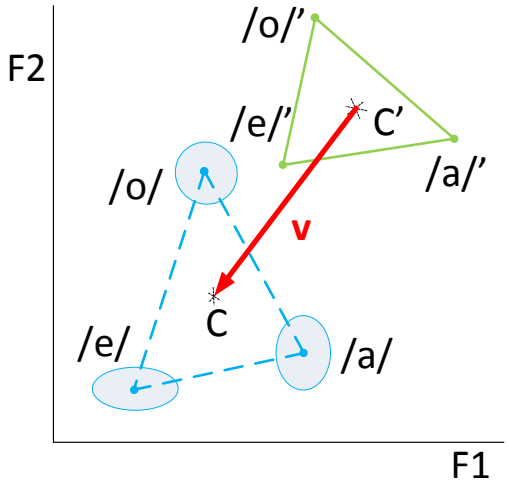

Fig. 9 The illustration of normalised and real small vowel triangle in the F1-F2 plane.

Experiments in speaker recognition used properties of vowel polygons were also used for finding the most variable feature within the speaker database. First steps of research were oriented on the observations of generated normalisation vector between reference and real vowel triangle, exactly between their centres of gravity [117]. Figure 9 illustrates the situation of normalisation vector **v** generation between reference (solid green line) and real (dashed blue line) small vowel triangle in F1-F2 plane.

TABLE V EXPERIMENTALLY ACHIEVED DIFFERENCES

Plane	**Big triangle**		**Small triangle**	
	Δd_{AVG} [Hz]	**$\Delta \alpha_{AVG}$ [°]**	**Δd_{AVG} [Hz]**	**$\Delta \alpha_{AVG}$ [°]**
F1-F2	17.5	17.7	16.0	45.2
F1-F3	30.7	23.1	31.1	36.3
F2-F3	28.1	65.2	**48.4**	**50.6**

In performed experiments, so-called big (AIU) and small (AEO) vowel triangle were used. Reference vowel triangles were generated by ordinal formant values of Czech vowels obtained by statistical measurements [111] and real vowel triangles were generated for each speaker by individual average formant values. For the created speaker database containing 12 male Czech speakers, the uniformity of created normalisation vectors were observed in the total number of three different formant planes (F1-F2, F1-F3 and F2-F3).

The speaker uniformity was observed in length and angle criterion of vector **v** by averaged absolute differences between each normalisation vector with average value of observed parameter over speaker database.

Table 5 contains experimentally achieved results. Obviously, the biggest differences in length (Δd_{AVG}) and angle ($\Delta \alpha_{AVG}$) criterion were reached for both vowel triangles in F2-F3 formant plane, but the best results were achieved by small vowel triangle. Figure 10 shows the polar distribution of real normalisation vectors **v** for both vowel triangles in formant plane F2-F3

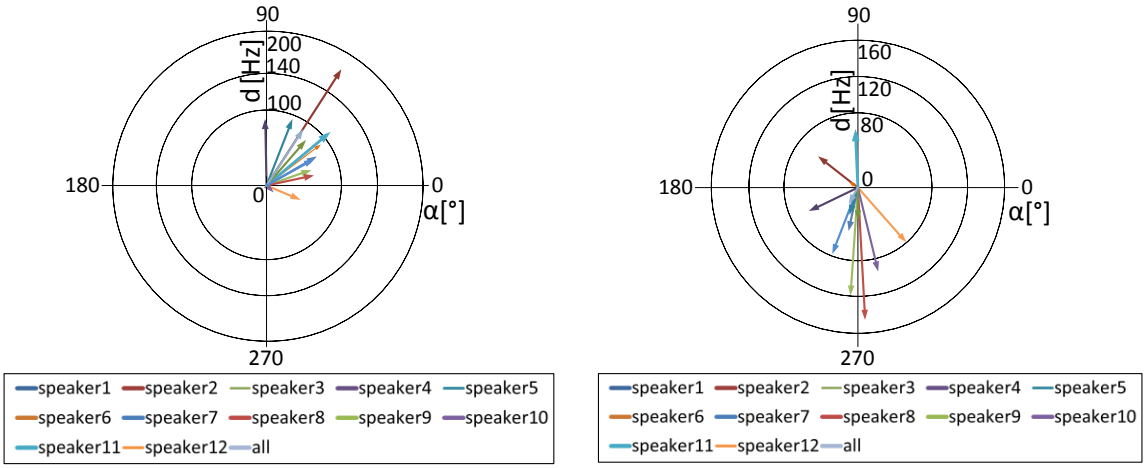

Fig. 10 Polar plots of vectors v for big (left) and small (right) vowel triangle in F2-F3 plane.

Furthermore, the uniformity of normalisation vector parameters was classified by statistical method ANOVA. Results reached by statistical testing are listed in Tab. 6. Generally, recognizing speakers by both parameters of normalisation vector created for small vowel triangle together is the best choice. All intermediate results of provided experiments, detail description and laid conclusion can be found in references [117].

TABLE VI RESULTS OF PARAMETERS TESTING USED ANOVA

Parameter	**Big triangle**		**Small triangle**	
	***F*-ratio [-]**	***p*-value [-]**	***F*-ratio [-]**	***p*-value [-]**
d	6.232	9e-5	2.821	0.016
α	2.031	0.071	0.521	0.870
***d, α* (Two-way ANOVA)**	3.089	6e-4	**8.371**	**7e-10**

TABLE VII AVERAGE FORMANT FREQUENCIES OF CZECH CARDINAL VOWELS USED FOR THE CALCULATION OF REFERENCE TRIANGLE AREAS

Formant	**Vowel**				
	$/a_{ref}/$	$/e_{ref}/$	$/i_{ref}/$	$/o_{ref}/$	$/u_{ref}/$
F1 [Hz]	900	590	400	600	400
F2 [Hz]	1300	1830	2400	1025	800
F3 [Hz]	2750	2750	3050	2750	2650
F4 [Hz]	3779	3403	3511	3974	3963
F5 [Hz]	4577	3909	3869	4923	4981

Developed software system [115] was used for achieving new results in the field of speaker recognition continuing on preliminary results [117]. The new observations were based on are differences between real and reference vowel triangles [118]. For created speaker database containing 13 male and 12 female Czech native speakers, reference values of higher formants F4 and F5 were achieved. Average Czech vowel formant values listed in Tab. 7 were used as the apexes of reference vowel triangles. Firstly, the measurement was focused on the estimation of the amount of vowel signal needed for calculating reliable formant parameters. Experimental results show that a data set of approximately 3000 values (i.e., 3000 vowel frames) satisfies statistical reliability. Figure 11 shows the development of cumulative values of mean F4-frequency calculated from the vowel /e/ of three male and three female speakers. The resulting trend corresponds to the steadying process of the fundamental frequency mean value published in [119].

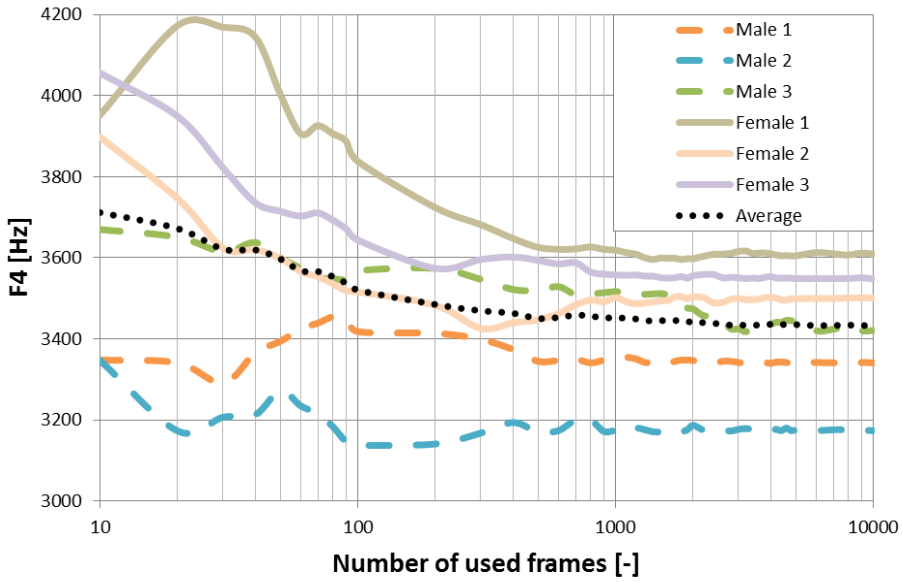

Fig. 11 Cumulative values of mean F4 obtained from five speakers in sets of 10 to 10000 speech frames of vowel /e/ (x-axis is logarithmic due to better resolution).

Further, research method based on relative differences between reference and real vowel triangle area was applied on another created speaker database consisted of 12 different male Czech native speakers. The total number of 10 vowel triangles can be created but in provided experiments were used only nine vowel triangles due to line character (zero area) of AEO vowel triangle in all planes containing the third formant which leads to infinite area differences. Owing to five formants can be possibly occurred in spectrum, experiments were provided for ten different formant planes.

The consistence of achieved area differences was classified by statistical indicator defined by following equation

$$R = \frac{\sigma_{dS}}{dS_{avg}}, \tag{6}$$

where σ_{dS} is a standard deviation of area differences of current vowel triangle and formant plane over the investigated speaker database, and R is called coefficient of variation. Due to including the standard deviation of actual selection, the consistence of area differences dS is established by R ratio for all speakers. The high suitability of actual vowel triangle, exactly of its area difference to reference pattern, is represented by the possibly lowest value

of *R* ratio. Vowel triangles are placed in order in Tab. 8 in the case of the coefficient of variation value, where triangles are titled by vowel created their apexes and two numbers sign formant plane.

TABLE VIII VOWEL POLYGONS RANKED BY ACHIEVED COEFFICIENT OF VARIATION

Order	Triangle	*R* [-]	Order	Triangle	R [-]
1.	AIO15	0.09	…		
2.	AEU15	0.11	86.	AOU24	1.37
3.	AIU15	0.13	87.	AIU45	1.38
4.	EOU15	0.14	88.	AIO23	1.40
5.	EIO25	0.17	89.	AIU35	1.49
…			90.	AOU45	2.03

The curves behaviour for AIO vowel triangle over all formant planes is illustrated in Fig. 12 where the current maximal area difference is performed by solid black line. Actual minimal area difference related to current formant plane is represented by black dotted line and the average area difference dS_{avg} is figured out by grey dashed line. Even the area differences are not high in comparison with other formant planes, the most suitable formant plane for speaker recognition using AIO vowel triangle is F1-F5 because of the lowest *R* value (see Fig. 12).

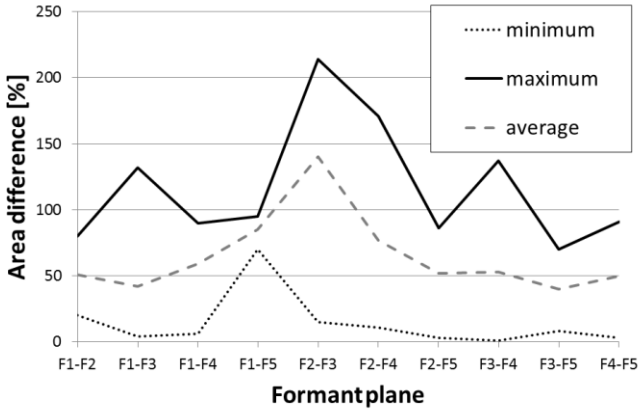

Fig. 12 Achieved area differences for AIO vowel triangle.

TABLE IX AVERAGE COEFFICIENT OF VARIATION FOR TOP FIVE VOWEL TRIANGLES AND FORMANT PLANES

Order	**1.**	**2.**	**3.**	**4.**	**5.**
Triangle	EIO	AEI	EIU	EOU	AEU
R_{avg}**[-]**	0.35	0.36	0.49	0.51	0.55
Plane	F1-F5	F1-F2	F1-F4	F3-F4	F3-F5
R_{avg}**[-]**	0.23	0.49	0.53	0.56	0.63

Final evaluation of triangle and plane speaker recognition suitability is listed in Tab. 9, where R_{avg} is an average coefficient of variation over relevant criterion. In the case of the lowest R_{avg} value, the best results are achieved by EIO triangle tightly followed by AEI triangle. For both mentioned triangles, the individual R_{avg} values are more a less equal. In plane criterion, the least R_{avg} value has been reached by formant plane F1-F5. For followed

formant plane, the second least R_{avg} value of F1-F2 formant plane achieved more than double the least value which signifies the rapid increase of relevant standard deviation of area differences. As it can be seen from Table 9, the triangle criterion is more applicable for speaker recognition as well as F1-F5 plane.

Obtained results and research method were presented at international conference and are described in more detail in relevant proceedings [118].

Further research oriented on speaker recognition is based on all possible vowel polygons (10 triangles, 4 tetragons and 1 pentagon) in ten different formant planes [120]. Exactly, the dispersion vector **d** is generated for each possible couple (two different speakers) of real Centres Of Gravity (COGs) and its significant length values are observed. Figure 13 shows the example of situation containing COGs of EIU12 as the expression of method idea. It has to be said, all possible vectors **d** are not shown in the case of better illustration.

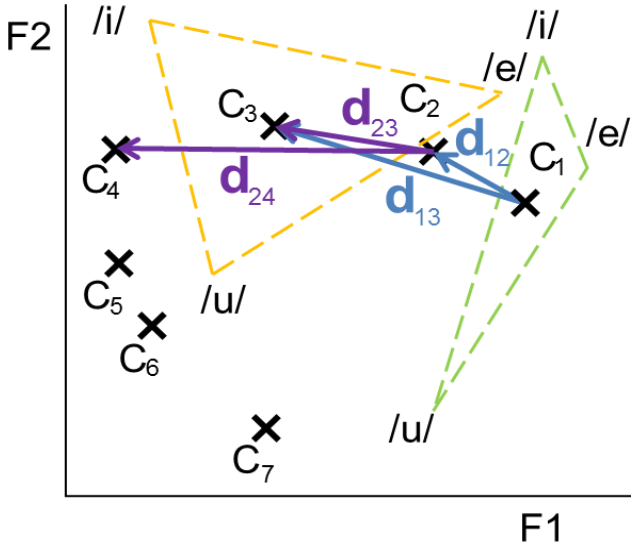

Fig. 13 The illustration of possible vectors d created by centroids of EIU12 vowel triangle.

The length of each vector **d** is calculated and then the length differences between all possible vector couples are observed. For the total number of speakers N, the minimal vector length difference is achieved from the following symmetric matrix

$$P_{diff} = \begin{pmatrix} |d_1 - d_1| & |d_1 - d_2| & \cdots & |d_1 - d_N| \\ |d_2 - d_1| & |d_2 - d_2| & \cdots & |d_2 - d_N| \\ \vdots & \vdots & \ddots & \vdots \\ |d_N - d_1| & |d_N - d_2| & \cdots & |d_N - d_N| \end{pmatrix}, \tag{7}$$

where all elements of the main diagonal are null. The desired parameter $\Delta d_{\min}$ representing the suitability of relevant vowel polygon for speaker recognition is defined as the minimal value of upper triangular part of matrix $\boldsymbol{P_{diff}}$. In the case of satisfied speaker recognition, the value of observed length differences $\Delta d_{\min}$ has to be maximized.

Presented method was applied on created speaker database consisted of fluent text spoken by 13 male and 12 female speakers and separately spoken vowels by another 12 male speakers. The list of vowel polygons ordered with respect to $\Delta d_{\min}$ criterion is shown in Tab. 10, where other important values of vector length difference can be found as the average Δd_{avg} and maximum value Δd_{max} over the created speaker database (37 speakers).

TABLE X THE LIST OF VOWEL POLYGONS ORDERED BY MINIMAL VECTOR DIFFERENCE VALUE

Order	Vowel Polygon	Δd_{min} [Hz]	Δd_{avg} [Hz]	Δd_{max} [Hz]
1.	EIO25	19	252	686
2.	AEU25	16	255	728
3.	EIOU25	16	247	586
4.	AIO35	15	257	773
5.	IOU34	14	215	639
6.	AEIO25	13	255	775
7.	AIOU35	13	243	592
8.	AIU45	12	260	706
9.	AOU45	12	240	696
10.	IOU23	12	213	610
…				
156.	IOU15	1	188	556
157.	AEOU35	1	247	627
158.	AEIU12	1	121	482
159.	EIOU13	1	144	458
160.	EIO13	0	148	436

The best minimal vector length difference is achieved by EIO vowel triangle in formant plane F2-F5, where the value of Δd_{min} approaches 19 Hz. The null Δd_{min} for EIO13 triangle is caused by the same COG coordinates of two female speakers, see Fig. 14. Obviously, the minimal values of Δd_{min} are achieved in formant planes containing the first formant F1 which sets together with F2 the spoken vowel. The distribution of COGs is illustrated in Fig. 14 for the EIO13 vowel triangle. The identical centroids causing the null value of Δd_{min} (see Tab. 10) are marked by purple circle. Figure 14 demonstrates the distribution of real COGs for EIO25 vowel triangle which reached the best value of Δd_{min}. In following figures, the real positions of COGs for female speakers are circles filled by green colour, male fluent speech is represented by red filled circles and separately spoken vowels by other male speakers are labelled by blue colour.

The suitability of vowel polygons for speaker recognition can be sorted and observed by another criterion called dispersion coefficient δ, which represents the relative variability ratio slightly influenced by extreme values and is defined by following equation

$$\delta = \frac{\frac{1}{N-1}\sum_{i=1}^{N}\left|\Delta d_i - \Delta d_{avg}\right|}{\Delta d_{avg}}, \tag{8}$$

where N is the total number of speakers.

Sorted values of dispersion coefficient δ in ascending order are listed in Tab. 11. The best vowel polygon suitability for speaker recognition is performed by the minimal δ value reached by EIO34 vowel triangle. Generally, the least speaker uniformity is reached in F3-F4 formant plane (see Tab. 11) in the case of δ criterion. Absolutely highest speaker uniformity has been reached in F1-F2 formant plane which proved the statement of vowel determination by the position of first two formants.

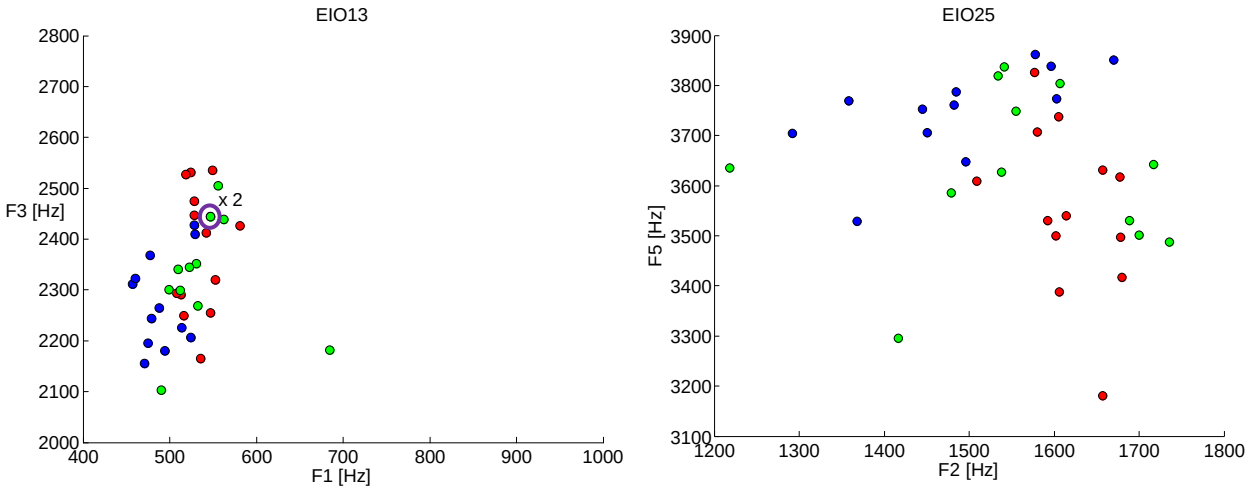

Fig. 14 The distribution of COGs positions for EIO13 (left) and EIO25 (right) vowel triangles.

TABLE XI THE LIST OF VOWEL POLYGONS ORDERED BY DISPERSION COEFFICIENT

Order	Vowel Polygon	δ [-]
1.	EIO34	0.39
2.	EIU34	0.40
3.	EIOU34	0.40
4.	AEU34	0.40
5.	AIU34	0.40
6.	AOU34	0.41
7.	EIOU25	0.41
8.	AEIU34	0.41
9.	IOU34	0.41
10.	AIO34	0.41
…		
156.	AIO12	0.60
157.	AIOU12	0.61
158.	AEOU12	0.61
159.	IOU12	0.66
160.	AEO12	0.66

The comparison of COGs distribution for the worst and the best vowel polygons is illustrated in Fig. 15.

By presented results, the final statements can be laid. The most suitable vowel polygons for speaker recognition are EIOU25 tetragon and IOU34 vowel triangle, which reached one of the least δ and one of the highest $\Delta d_{\min}$ values. Generally, the most suitable shape for speaker recognition is the vowel triangle and the least speaker uniformity has been reached in higher formant planes, while the absolutely unsuitable formant plane for speaker recognition is F1-F2 determining especially spoken vowel. All presented results of COGs distribution were published in [120].

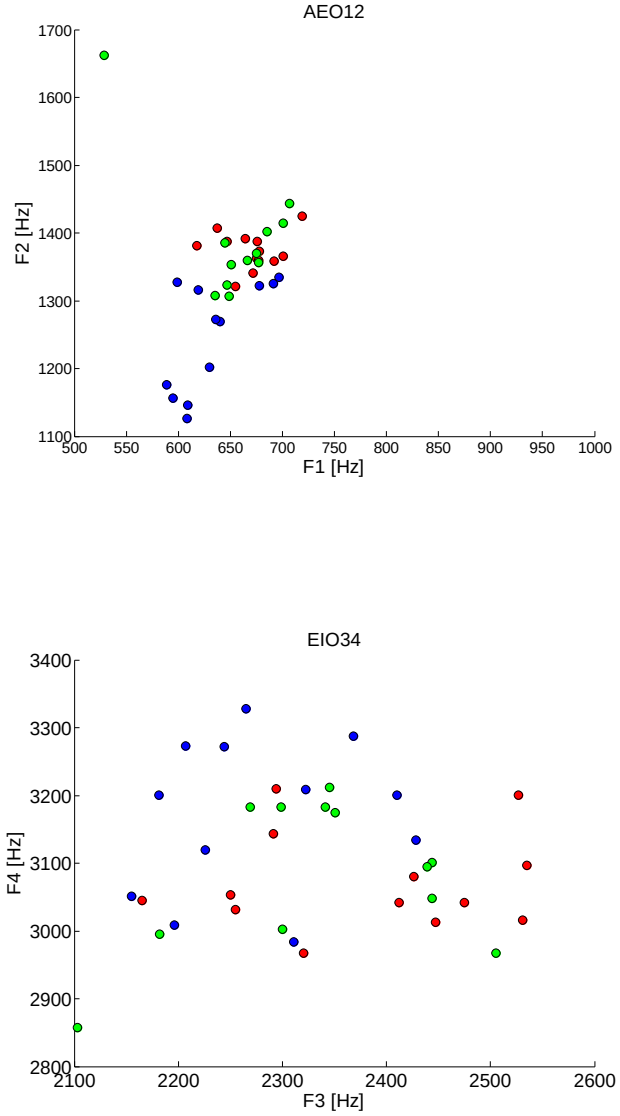

Fig. 15 The distribution of COGs positions for AEO12 (top) and EIO34 (bottom) vowel triangles.

6.3. Psychological stress and vowel polygons

All experiments oriented on observing the differences between normal speech and speech under psychological stress influence were provided on special created database. The first idea of this created database was presented in [121]. Basically, the more a less same text is spoken by each speaker in normal mood and in situation causing stress influence. For introducing observations of psychological stress influence in speech, the total number of 18 Czech male speakers was recorded during the final exam and the defence of master's thesis where the stress influence is assumed.

Primary observations were also based on properties of vowel polygons, exactly if properties of vowel polygons generated from stressed speech tend to same behaviour. In primary experiments, this prediction was confirmed. Firstly, the differences of vowel polygons COGs positions between normal mood and stress influence were observed. Figure 16 shows created by normal mood COGs (the vector beginning) and stress influence COGs (the vector end- arrow) for 18 male speakers and AIU vowel triangle in formant plane F2-F3.

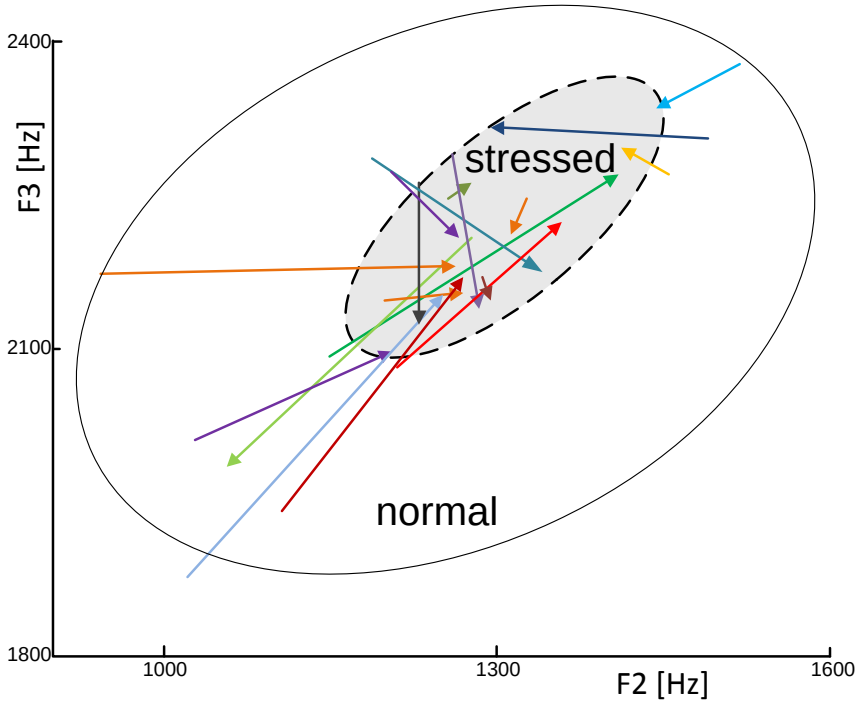

Fig. 16 Differences of CGGs positions for normal and stressed speech in AIU23 vowel triangle.

In Fig. 16, the changes of COGs concentration are evident. Normal mood boundary is illustrated by solid line ellipse and stress influence is performed by dashed line ellipse filled by grey shadow. The dimensions of ellipses are defined by relevant COGs. Obviously, the area of normal mood ellipse is several times bigger than the area of stress influenced ellipse which means the COGs concentration is much higher for stress speech. For speech under stress, it is typical the ellipse rotation to vertical axis (higher formant axis) as well as the much

smaller minor axis dimension than the major axis which leads to focal shift closer to the edge. These findings were occurred for all possible vowel triangles in all formant planes.
Due to this uniform behaviour of vowel polygons under psychological stress, further experiments were provided to make closure of vowel polygons part. Following results were published in [122, 123] and applied on previously described database ExamStress [121], exactly on randomly selected 18 male Czech native speakers telling the same text during and after final exam, which means two identical records differed only in emotion state are received for each speaker. Stress contained in used database is divided into three groups- low, middle and high stress influence. Only middle and high stress records supplementing their normal speech equivalents are used in performed experiments.

These records represent the input of developed and further used software system generating and analysing vowel polygons [115]. Briefly, each input sound record is resampled to f_s= 8kHz, and further vowels are recognized from fluent speech by using two-level recognition system (Mahalanobis distance, Forward-feed Neural Network), retroactively checked [113] and the values of all occupied formant frequencies in each vowel are saved for further processing. In the case of used sampling frequency, at most five formants can be occurred in LPC spectrum, which lead to the total number of ten possible formant planes. As it was mentioned, presented research is oriented on Czech language containing five vowels /a/, /e/, /i/, /o/, /u/ and their so-called long equivalents differing only in duration not in pronunciation. The total number of five Czech vowels leads to sixteen different shapes (ten triangles, five tetragons and one pentagon) which can be investigated. These shapes situated in formant planes are called vowel polygons and their generation, marking and other information can be found in [118].

6.3.1. Stress division

Obvious signs of vowel polygon behaviour depending on normal and stressed state of speaker are observed in two criteria. Firstly for each vowel polygon, the area differences between actual (stress) and original (normal) are observed for investigating the possible uniform behaviour of this parameter as well as the direction and length of vector facing from original to actual Centre Of Gravity (COG). Figure 17 shows generated vectors for AEI vowel triangle observed in formant plane F3-F4, where each individual ellipse presents middle stress, high stress and both stress states vectors (from top to bottom).

For the majority of all possible vowel polygons, same effects are occurred as well as for illustrated example (see Fig. 17). Firstly, created vectors are more uniform in their direction for high stress influence, and their angle reaches approximately value $\pm \pi/4$. The direction of middle stress vectors is more different, higher deviation between individual vectors, but they are still pointing similarly to high stress influence. Generally, stress influenced vectors are not occupied in the second and fourth quadrant. By these statements and previous research [117], the increasing direction uniformity of created vectors can be assumed with increasing stress level which leads to erasing the deviations between speakers.

Following observations are focused on getting the cross-correlation values between vowel polygon area difference and one parameter of created vector. These values are also further statistically analysed by coefficient of variation (equivalent to equation 6) defined as follows

$$R = \left| \frac{\sigma_x}{\bar{x}} \right|, \tag{9}$$

where σ_x is standard deviation of observed parameter x (e.g. cross-correlation values of selected vowel shape over all formant planes) and $\bar{x}$ is its mean value. This statistical pointer

shows higher uniformity of received results by lower number leading to more reliable and significant results [124].

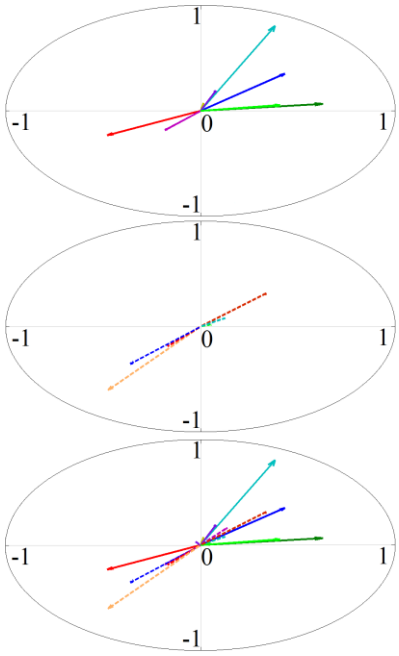

Fig. 17 Differences between AEI34 created vectors' length and direction for middle (top), high (middle) and mixed (bottom) stress level of 18 different speakers.

6.3.2. Method application

Following section is divided into three parts oriented on three different stress types- middle, high and mixed stress. As it has been uncovered in previous section, different stress types have to be analysed and observed separately for giving significant proofs because higher stress level leads to higher speaker uniformity in created vectors. The core of provided experiments is created by cross-correlation of chosen vowel polygon's parameters couples for achievement obvious relations between them. Nowadays, the cross-correlation is ordinary used in the speech processing in the field of emotion recognition [125], speaker [126] and speech [127] identification. Following results are colour distinguished to sign different couple of cross-correlated parameters. For simplification in following text, these colours also represent used experimental methods. Green colour signs the cross-correlation of difference area value and vector length (method 1), light blue represents signum of area difference and vector length (method 2), dark blue difference area value and vector angle (method 3), orange signum of area difference and vector angle (method 4).

The last two methods can be called mean methods, because of the usage of previously reached results received by two together related methods. Method 5 (purple) is defined as the mean of green and dark blue progressions. Light brown represents method 6- the mean of light blue and orange values.

6.3.3. Middle stress – experimental results

Used stress-influenced records were spoken by doctoral students and captured before trying to pass rigorous exam in front of commission. Normal state of speaker was recorded few days after the examination. Same text, exactly speech, is contained in both records and is spoken by 8 different male Czech native speakers. For all possible formant values, the correlation values over all shapes were calculated to observe general best formant plane and classification method. The boxplot of mined middle-stress data is shown in Fig. 18 where the median value of current data set is illustrated by red horizontal line, 25 and 75% percentiles are illustrated by coloured bar, whiskers show the residual current data set and outliers are marked by red crosses. Obviously, boxplots are used due to good representation of reached results for finding relationships between observed speech parameters for instance dyslexia [128], emotion [129] recognition, etc.

The colour coding in following figures differs reached results by used methods. As it was mentioned above, method 1 is represented by green colour, method 2 by light blue colour, method 3 by dark blue, method 4 by orange colour, method 5 by purple and the light brown colour stands for method 6.

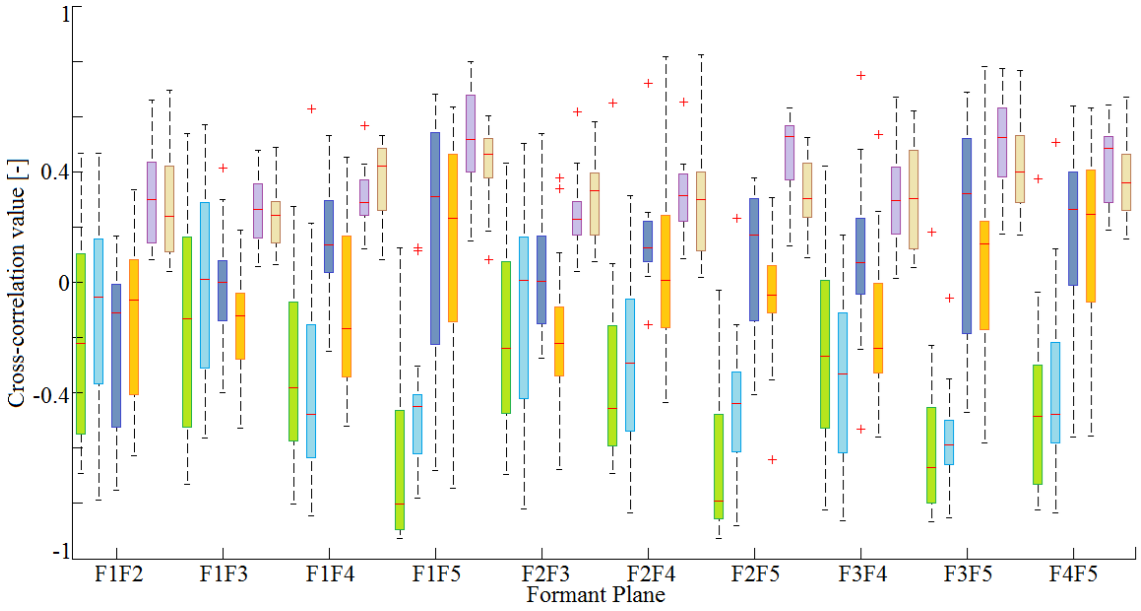

Fig. 18 Cross-correlation values over all vowel shapes for all used methods applied on middle stressed speech.

By the look at Fig. 18, few statements can be laid. The highest absolute correlation values are reached by both mean values (purple and light brown) and by cross-correlation of area difference and vector length (green). Generally, best correlation values are reached by green bars (area difference and vector length) and obviously between those two parameters exists a strong downhill (negative) linear relationship [130]. On the other hand, purple and light brown bars show only more a less moderate uphill (positive) relationships similar to light blue bars with inverse trend. Due to this fact, green bars better exhibit relationship between observed parameters. Obviously, area difference and vector length are the most correlated parameters in formant plane criterion for middle stress level. The uniformity of received correlation over formant planes is illustrated by *R* in Fig. 19 where the highest uniformity of correlation means

(purple and light brown) is evident. These two methods are significant by the least changing correlation values over all possible vowel polygons for each formant plane and the most uniform formant plane is generated by formants F2F5.

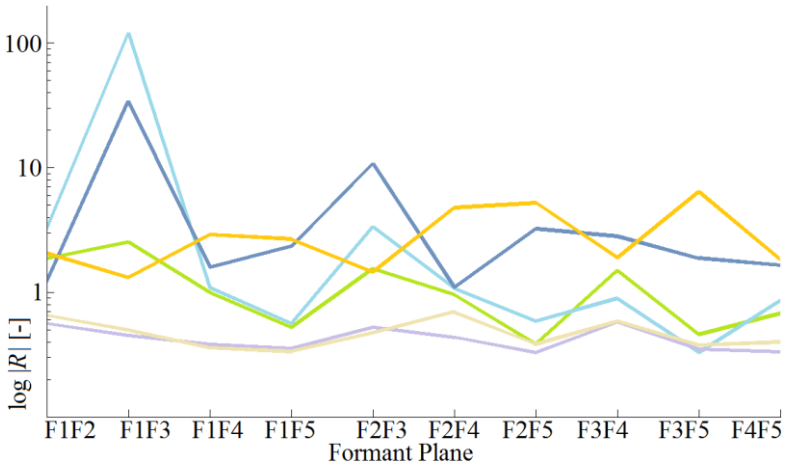

Fig. 19 Reached coefficient of variation in formant planes criterion applied on middle stressed speech (y-axis is in the logarithmic scale for better resolution).

Similar dependency of *R* on vowel polygons over all possible formant planes is shown in Fig. 20, where results similar to previously laid are presented. Again, both mean methods (purple and light brown) are characterized by lowest *R* values leading to the highest uniformity of correlation results. Most consistent cross-correlation results are received for AEIO vowel tetragon by purple mean method (see Fig. 20). Figure 21 illustrates the total plane distribution of used methods uniformity.

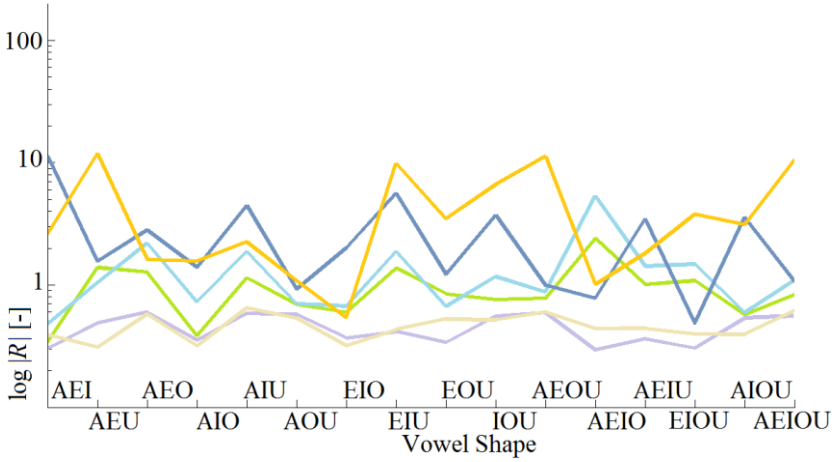

Fig. 20 Reached coefficient of variation in vowel shapes plane criterion applied on middle stressed speech (y-axis is in the logarithmic scale for better resolution).

Obviously, results achieved by mean methods (purple and light brown) are most concentrated and characterized by lowest *R* values, but the highest correlation results variability over shapes and formant planes is received by cross-correlation using signum of area difference and vector length (light blue) as well as cross-correlation of area difference and vector angle (dark blue). These two methods are absolutely unsuitable for stress detection because of the high cross-correlation value variability leading to insignificant proofs.

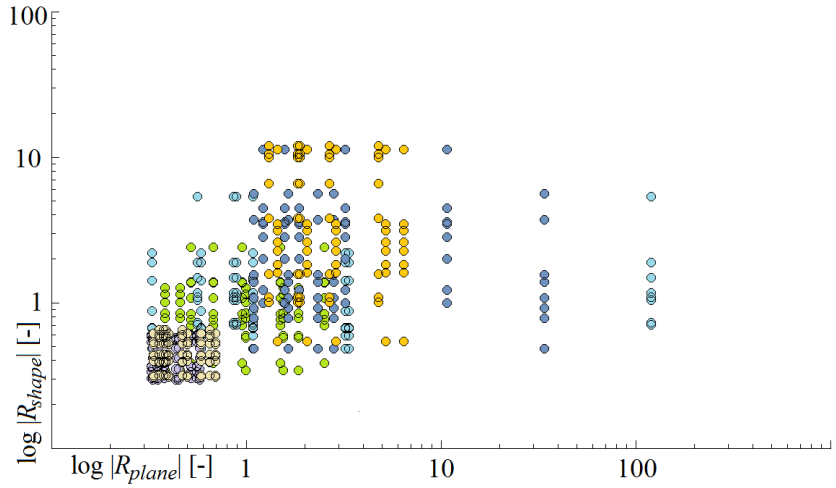

Fig. 21 Plane figuring out reached *R* for middle stress influence. The highest results uniformity is evident for both mean methods (purple and light brown).

6.3.4. High stress – experimental results

Comparing to previous stress level associated with rigorous exam in front of exam commission, high stress influence is caused by possible option which can lead to striking failure of current situation. Due to this reason, the stressor's pressure is more intensive on observe subject than in previous case [131].

Generally, results and processes presented in this subsection are similar to previously mentioned in subsection oriented on middle stress with the difference that records are captured for another 10 male Czech native speakers during and after master thesis defence faced to committee board. Due to this fact, higher stress level is presumed opposite to previous subsection and part of used database. Figure 22 presents boxplot containing cross-correlation values for six couples of observed parameters over all possible shapes for each plane. The marking and meaning are same as for Fig. 18.

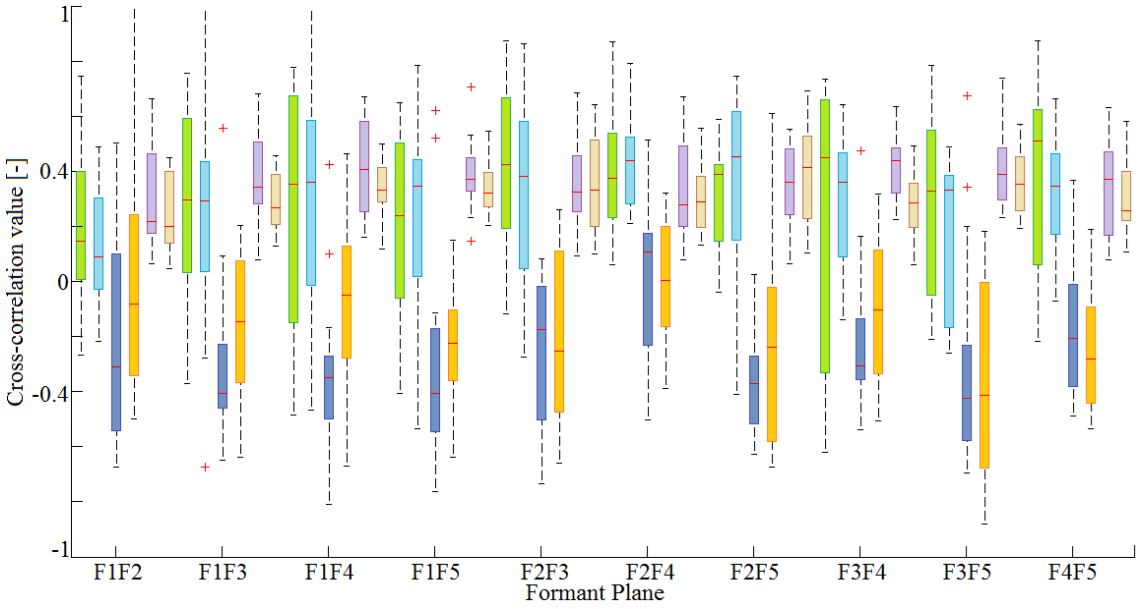

Fig. 22 Cross-correlation values over all vowel shapes for all used methods applied on high stress level.

Figure 22 shows that for high stress influence the cross-correlation value is generally lower than for middle stress level over each formant plane, which means the relationships between observed parameters couples are more individual within selected speakers. All received results present only weak down or uphill linear relationships between observed parameters. Obviously, Fig. 22 illustrates that the lowest *R* values are received also for both mean methods (purple and light brown).

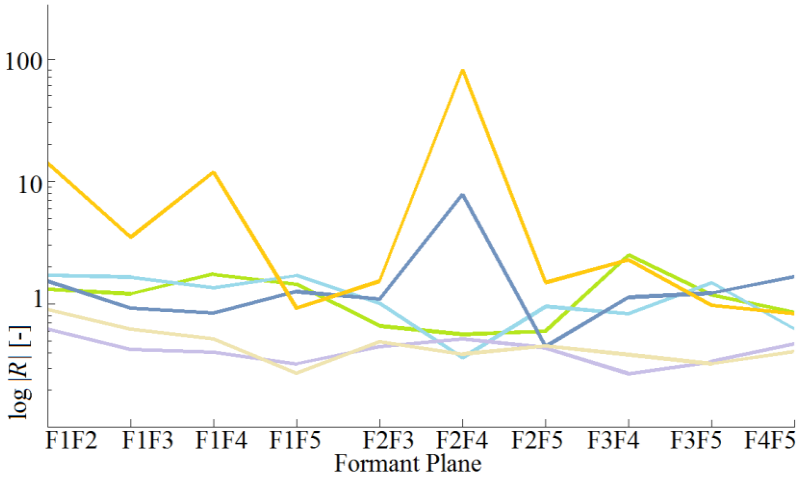

Fig. 23 High stress level- reached coefficient of variation in formant planes criterion (y-axis is in the logarithmic scale for better resolution).

The uniformity curves are shown in Fig. 23, where best result over vowel shapes is received for purple mean method in formant plane F3F4 and the worst, lowest achieved results uniformity, is reached by cross-correlation of area difference signum and vector angle (orange) in formant plane F2F4. According to similar results in previous subsection, the basic usage of vector angle is seemed useless for stress detection. Both mean methods reach higher uniformity of previously mined results by cross-correlation. This fact can be caused by the event where each subject feels more a less same stress level as the other caused by higher probability of final exam failure which leads to less self-confidence of each individual speaker and higher differences between normal and stressed speech. By comparison with Fig. 19, the uniformity of all used methods is higher and approaches more constant values than for middle stress level presented previously.

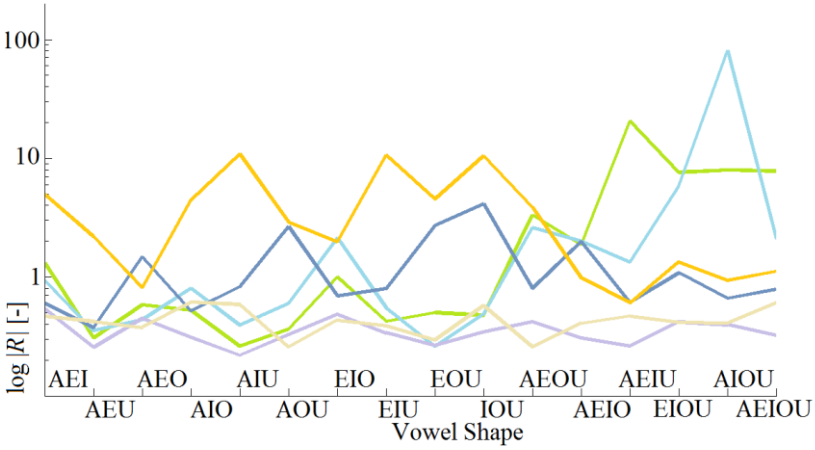

Fig. 24 High stress level- reached coefficient of variation in vowel shapes criterion (y-axis is in the logarithmic scale for better resolution).

The uniformity of achieved correlation results in shapes criterion over all formant planes is illustrated in Fig. 24. Figure 24 presents the highest uniformity is in shape criterion also reached for both mean methods (purple and light brown curves), but generally the uniformity of cross-correlation results is much better for higher stress level than for middle stress (comparison of Fig. 20 and Fig. 24). The lowest *R* value is reached by purple curve for AIU vowel triangle, the worst result uniformity is achieved for AIOU vowel tetragon by light blue curve (cross-correlation of area difference signum and vector length).

The consistency of mined *R* values is shown in Fig. 25, where the worst methods for stress detection are light blue (area difference signum and vector length) and orange (area difference signum and vector angle). By this observation, it can be set the statement of the unsuitability of using the area difference signum for higher stress detection leading to high cross-correlation results variability and insignificant high stress detection. On the other hand, the most uniform cross-correlation results are again received for both mean methods (purple and light brown) in plane and shape criteria. By this distribution illustrated in Fig. 25, both mean methods have been confirmed again as the methods reaching the most consistent results in formant plane and selected shape criterion.

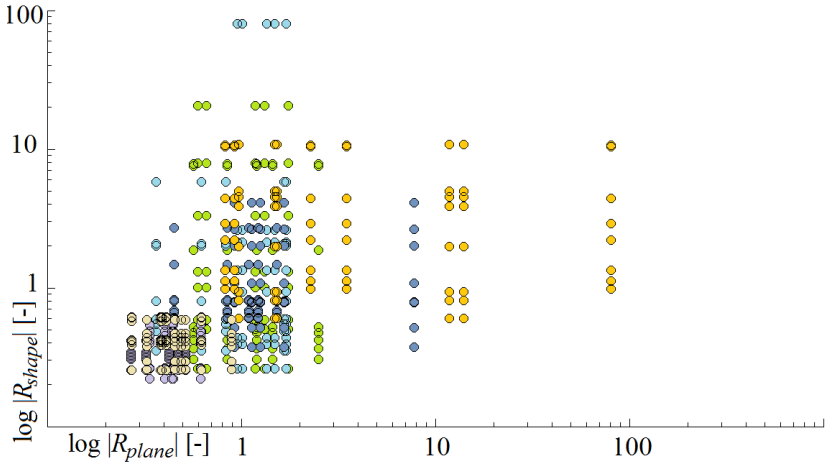

Fig. 25 Plane figuring out reached R for high stress influence. All points are more concentrated together than in previous case, see Fig. 5.

6.3.5. Mixed stress – experimental results

Previously uncovered results obviously lead to the achievement of higher cross-correlation values for middle stress level influence. Cross-correlation results received for middle stress level are also more consistent, or uniform, than for higher stress influence which slightly disproves the original idea based in Fig. 17 of increasing cross-correlation values relating to higher stress level. Following results presented in this subsection are received by mixing both previously used partial databases of middle and high stress level influence. The boxplot of experimentally achieved cross-correlation coefficient values over all possible shapes is illustrated in Fig. 26. Obviously, the worst results of cross-correlation values uniformity are received currently for mixed stress states.

As in previous findings, the highest uniformity of received results in the plane criterion is received for both mean methods (purple and light brown) leading to following figure, Fig. 27, where a huge difference between mean and other methods is evident and moreover, the *R* reaches higher and less suitable values for other methods (green, light blue, dark blue, orange) similarly to *R* in over shapes presented in Fig. 28, which signs high possible failure of

methods 1-4 if they are chosen to stress detection. By these two presented figures (Fig. 27 and Fig. 28) and previously introduced findings, the high level separation of mean methods from the others is expected in plane expression of *R* in plane and shape criterion.

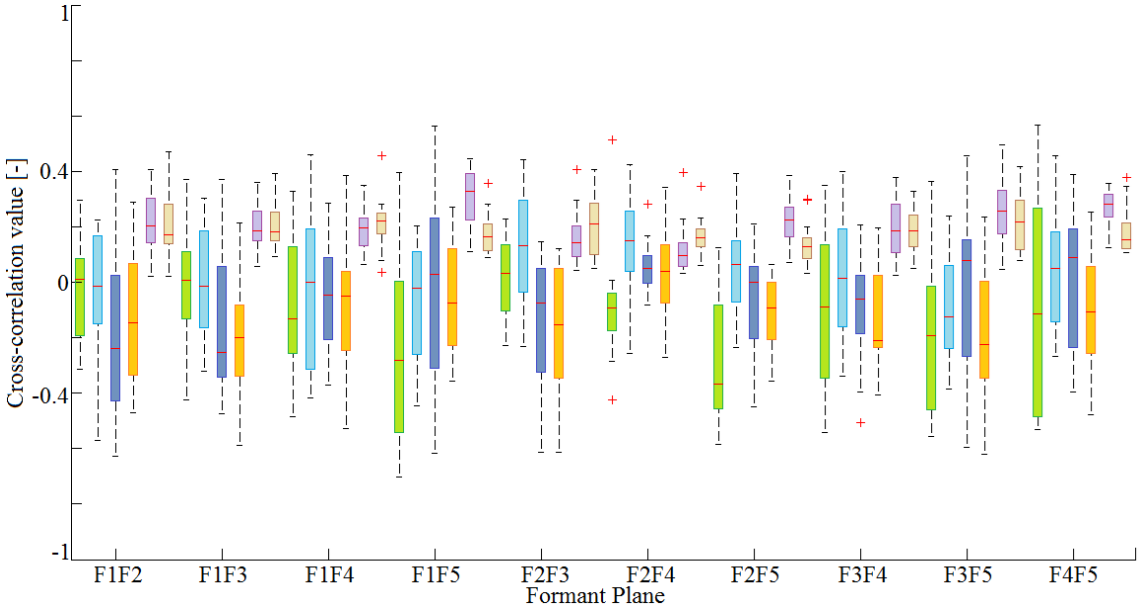

Fig. 26 Cross-correlation values over all vowel shapes for all used methods applied on mixed (middle + high) stress influence.

As the final presentation of partial results (see Fig. 27 and Fig. 28), Figure 29 is created. Figure 29 shows mentioned huge uniformity difference between mean and other methods for plane as well as shape criterion. Generally for mixed stress, *R* in both criteria reaches higher value than in previous cases, but mean methods (purple and light brown) are more uniform which leads to their absolute separation from other methods (see Fig. 29).

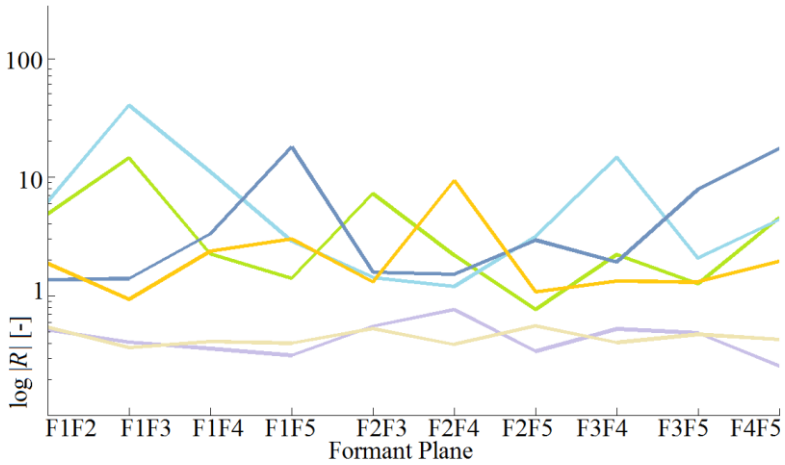

Fig. 27 Reached *R* in formant planes criterion applied on mixed stressed speech (logarithmic y-axis).

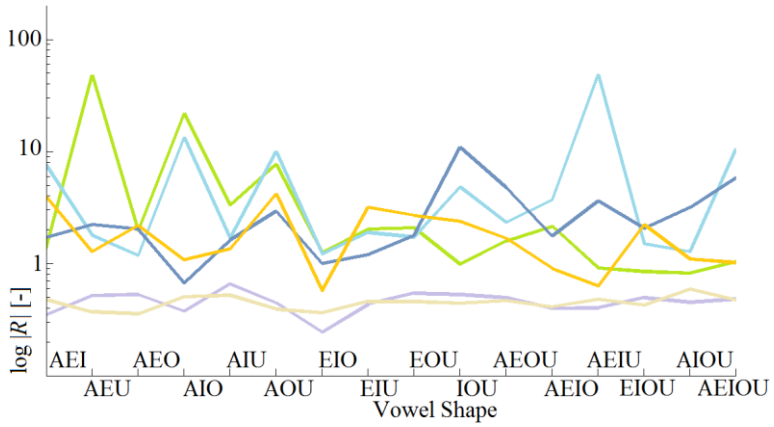

Fig. 28 Reached *R* in vowel shapes criterion applied on mixed stressed speech (logarithmic y-axis).

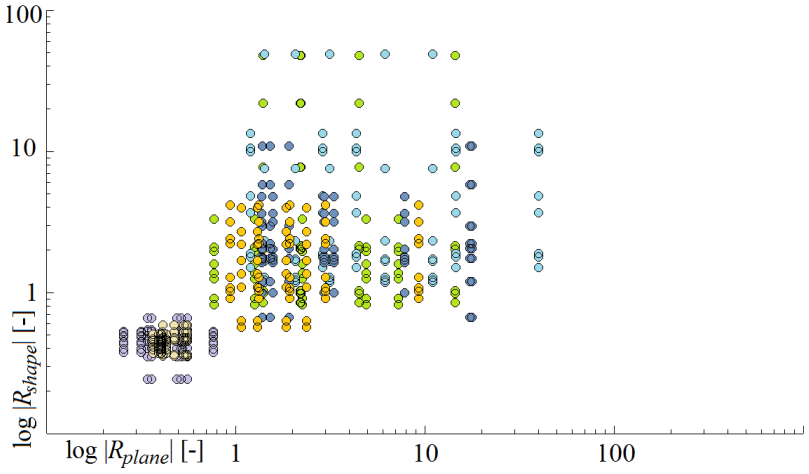

Fig. 29 Plane figuring out reached *R* for mixed stress influence. Both mean methods (purple and light brown) are absolutely separated from the others due to high results uniformity.

Recently introduced results can be partially summarized by following observations. The direction and length of generated vector by normal and stressed COGs are more equal with raising stress level within the speaker database. But the movement of observed couples of parameters is more a less individual for each speaker with higher stress level, with decreasing stress influence the cross-correlation of area difference and vector length or its direction is higher within speakers.

Generally, the most consist and significant differences between normal speech and stress influence are reached by both mean methods (purple and light brown) in shape as well as in the plane criterion. The advantage of these both mean methods significantly increases with higher stress level.

6.3.6. Efficiency of vowel polygons

In following sections, the suitability of stress detection will be observed for each possible vowel polygon separately because of not so significant results were achieved only in separated shape or plane criterion. The suitability, exactly the most significant and consist differences, are classified by their current efficiency which is based on results presented in previous section. Generally, the efficiency of observed parameter x is defined by equation

$$\varepsilon = \frac{\sigma_x^2}{\bar{x}^2}, \tag{10}$$

which can be modified in our case of usage as following equation of efficiency coefficient E_c

$$E_c = \frac{CCV^2}{\left|R_{shape}\right| \cdot \left|R_{plane}\right|}, \tag{11}$$

where CCV is previously calculated cross-correlation value for selected couple of observed parameters for current vowel polygon, R_{plane} is variation coefficient of relevant formant plane and R_{shape} of shape. Briefly, the value of efficiency coefficients signs the strength of observed couple of parameters for actual vowel polygon referred to statistical values over all relevant planes and shapes. The strength of observed vowel polygon is directly proportional to the E_c value- with increasing E_c rises the impact of current vowel polygon over others similar and relevant.

As it has been mentioned, the suitability of each individual vowel polygon usage for stress detection is classified by so-called efficiency coefficient E_c which signs the strength of current vowel polygon parameters over all relevant formant planes and shapes.

Figure 30 illustrates the E_c values (z-axis) depending on selected formant plane (x-axis) and shape (y-axis) for middle stress level. Due to better results representation, the z-axis is in logarithmic scale.

The top subplot presents results achieved for area difference and vector length, previously known by green curve as method 1, centre plot presents cross-correlation between area difference and angle (dark blue, method 3) and bottom plot illustrates mean of cross-correlation values (purple, method 5) for area difference - vector length and area difference – vector angle couples used for relevant E_c calculation.

Each individual E_c value is projected to the coloured area under dependency plot. Under the dependency plot two lines are also occurred. The red dashed line links best E_c values reached by vowel shapes in different formant planes, e.g. the best E_c values are received in F1F5, F2F5 and F3F5 formant planes for EIOU tetragon by mean method (see Fig. 30, bottom subplot). These peaks are also signalized by red marks in dependency plot. The best E_c values reached by each individual formant shape over all formant planes are figured out by blue dashed line and by blue circle marks.

Obviously, the best E_c values are reached by mean method where almost all values are higher than 1 conversely to area difference – vector angle couple where all values are lower than 1 which signs low impact of chosen parameters couple for stress detection.

Figure 31 shows middle stress results using signum of area difference added to vector length (top subplot, previously known as light blue or method 2) and vector angle (centre subplot, previously orange colour and method 4). Bottom subplot illustrates efficiency coefficient calculated from mean of relevant cross-correlation values, previously known by light brown colour as method 6. Similarly to previous three dimensional plots and presented results, the lowest suitability for stress detection is reached by methods based only on vector angle cross-correlated with another parameter (e.g. E_c= 0.273 and E_c= 0.220), but the total efficiency increases significantly by averaging of both methods using vector angle leading to method 6 (E_c= 4.448, see Fig. 30 and Fig. 31).

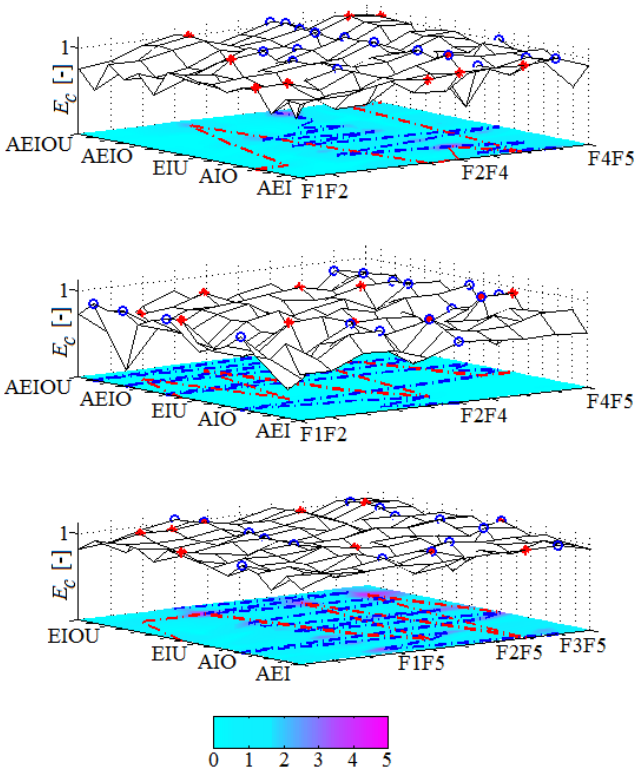

Fig. 30 Experimentally obtained E_c values for middle stress influence by methods 1, 3 and 5 (from top to bottom).

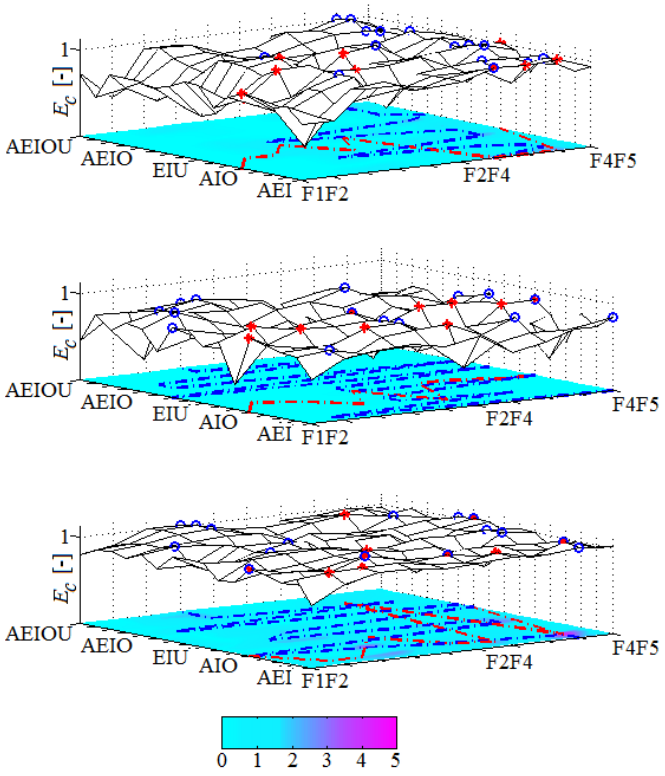

Fig. 31 Experimentally obtained E_c values for middle stress influence by methods 2, 4 and 6 (from to bottom).

By comparison of Fig. 30 and 31, not so significant final values differences exist between usage of area difference value and its signum. Concretely, maximum E_c values of each subplot for Fig. 30 (from top to bottom) are 4.149, 0.273 and 4.350 compared to Fig. 31 where peak value for each subplot (from top to bottom) is 2.697, 0.220 and 4.448. Other concrete E_c values are presented in following section.

According to maximal E_c values reached by vector length cross-correlated with other parameter, not so significant difference are evident between each individual results of method 1 and method 3 with their mean method , method 5, as well as for the usage of vector angle. But also the fact validates, the E_c value can be increased to higher value by averaging of mined results by method 1 and method 3 (e.g. E_c= 4.149 and E_c= 2.697, see Fig. 30 and Fig. 31) leading to generation of method 5 (E_c= 4.350). By this observation, total relationships between vowel polygon area difference, its signum, vector angle and length can be found by using suitable methods which can possibly lead to increasing the E_c for the most of vowel polygons.

Comparing to previous stress level associated with rigorous exam in front of exam commission, high stress influence is caused by possible option which can lead to striking failure of current situation. Due to this reason, the stressor's pressure is more intensive on observe subject than in previous case [131].

Generally, results and processes presented in this subsection are similar to previously mentioned in subsection oriented on middle stress with the difference that records are captured for another 10 male Czech native speakers during and after master thesis defence faced to committee board. Due to this fact, higher stress level is presumed opposite to previous subsection and part of used database. Figure 22 presents boxplot containing cross-

correlation values for six couples of observed parameters over all possible shapes for each plane. The marking and meaning are same as for Figure 18.

6.3.7. Summarization

This part presents experimentally achieved values of efficiency coefficient E_c for each vowel polygon for 3 different stress groups and 6 different observation methods. Due to big amount of achieved results, following tables list only the top and the bottom 5 values for each stress group and used method.

TABLE XII THE BEST AND THE WORST VOWEL POLYGONS CLASSIFIED BY EACH USED METHOD APPLIED ON MIDDLE STRESS

E_c [-]; Middle Stress						
Rank	**Method**					
	1	**2**	**3**	**4**	**5**	**6**
1.	AIO25 4.149	AEI25 2.697	AOU35 0.273	EIO23 0.220	EIOU15 4.350	AEU35 4.448
2.	AIOU25 3.618	AEI35 2.605	AOU45 0.266	EIO34 0.164	EIOU35 4.242	AIO15 3.344
3.	AIOU35 2.710	AEOU 2.525	EIOU35 0.251	AIO23 0.163	AEU15 3.727	AOU35 2.904
4.	EOU25 2.651	AOU35 2.448	EOU45 0.202	EIO25 0.146	EIOU25 3.521	AEU15 2.707
5.	AOU25 2.629	AIOU35 2.221	EIOU15 0.199	AIU23 0.140	EIO45 3.379	AEI35 2.548
…						
156.	AEIO25 0.001	AEOU13 0.000	AIU35 0.000	AEO35 0.000	AEO12 0.019	AIU24 0.009
157.	AOU34 0.000	AEIU34 0.000	AIU15 0.000	AEIOU24 0.000	AIOU13 0.014	AIU34 0.008
158.	AIOU12 0.000	AEIOU13 0.000	AEO13 0.000	AOU12 0.000	AIU34 0.011	AEI12 0.006
159.	AIU34 0.000	IOU13 0.000	AIO13 0.000	EIOU13 0.000	IOU23 0.005	IOU24 0.006
160.	IOU23 0.000	AEO13 0.000	AIOU13 0.000	IOU25 0.000	AOU34 0.000	AEIOU24 0.000

Table 13 is equivalent to Tab. 12 with the difference of listed results are connected to high stress influence. Obviously, listed efficiency coefficients reach lower values than for middle stress level, but also the significant results are obtained. Best results are achieved also for method 5 followed by method 6 and method 1. The difference between method 5 for middle and high stress influence is not so striking but locally E_c values for high stress level are better. As in the previous case, the worst results are reached by method using cross-correlation of vector angle with area difference and its signum.

An interesting fact by comparison Tab. 12 and Tab. 13. is also the bottom of list (the worst values), where significantly higher E_c values can be found for higher stress level. From mined results, the best shapes are AIU, AEU and AIO vowel triangles, supplemented also by formant planes F1F5, F2F5 and F3F5.

By conjunction of middle and high stress partial database, results of mixed stress are achieved and listed in Tab. 14. Due to previously presented result differences between middle and high stress level separately, the decrease of reached E_c value is presumed for each testing method.

TABLE XIII HIGH STRESS INFLUENCE- THE LIST OF THE TOP AND BOTTOM VOWEL POLYGONS

E_c [-]; High Stress						
Rank	**Method**					
	1	**2**	**3**	**4**	**5**	**6**
1.	AIU23 1.920	AEU12 1.661	AIO14 1.373	AEIO35 0.228	AEU23 4.320	IOU13 2.803
2.	AIU14 1.911	AIU12 1.491	AEU23 1.223	AEIU13 0.193	AIO35 4.141	AEU12 2.655
3.	AIU13 1.807	EOU24 1.450	AIO15 1.222	AEO35 0.186	EOU24 3.998	AIO14 2.619
4.	AOU23 1.754	AEU23 1.269	AIO35 1.008	AEIU25 0.184	AIU14 3.833	AIU12 1.910
5.	AOU24 1.735	IOU13 1.234	AIO23 0.982	AIOU35 0.157	AIU13 3.823	AEU25 1.827
…						
156.	AEIU35 0.000	AEIOU13 0.000	EIU23 0.002	AIU14 0.000	AEO24 0.031	EIU12 0.030
157.	AEIOU13 0.000	AIOU13 0.000	IOU23 0.002	AEO24 0.000	AEI13 0.026	AEIOU45 0.030
158.	AEIU12 0.000	AIOU34 0.000	AEO24 0.002	EIU25 0.000	AEI12 0.026	AIO25 0.028
159.	AIOU35 0.000	AIOU12 0.000	AIOU25 0.001	AIU24 0.000	EIO25 0.020	AEIOU34 0.010
160.	AIOU12 0.000	AIOU45 0.000	EOU23 0.000	AEOU13 0.000	AEO12 0.015	AEIU12 0.005

As it has been presumed, the significant general decrease of all E_c values is obvious and leading to statement the only applicable method to mixed stress detection is method 5 using the mean of area difference value – vector length and area difference value – vector angle cross-correlations. Despite the adverse E_c decrease, the fifth method reaches values higher than some others for partial stress classification.

Similarly to previous cases, the best results are achieved by method 1, method 5 and method 6 with the exception that only method 5 can be further usable because some of its E_c values reach values higher than 1. The general best shape choices are AIO, EIO vowel triangles and AEIO vowel tetragon for mixed stress classification as well as formant planes F1F5, F2F5 and F2F3. Absolutely best result is achieved by AIO15 vowel triangle reaching the E_c value equivalent to the circa sixth-tenth place of Tab. 12 and Tab. 13. By results listed

in this section, the usage of vowel triangles and formant planes containing the formant F5 can be finally evaluated as the best choices for stress detection.

TABLE XIV THE BEST AND THE WORST VOWEL POLYGONS ACHIEVED BY EACH USED METHOD FOR MIXED STRESS LEVEL

E_c [-]; Mixed Stress						
Rank	**Method**					
	1	**2**	**3**	**4**	**5**	**6**
1.	AIOU25 0.498	EOU23 0.080	AIO23 0.357	AEIU13 0.592	AIO15 1.612	AEO45 0.937
2.	EIOU15 0.415	EOU24 0.058	AIU12 0.129	AEIO35 0.325	AEOU15 1.266	AEIO35 0.882
3.	EIOU25 0.312	AEU23 0.056	AIO34 0.123	EIO23 0.269	AEIO45 1.255	AEIU13 0.864
4.	AEIU25 0.311	AIOU15 0.054	EOU12 0.118	AIO23 0.268	AIO35 1.246	AIOU14 0.851
5.	EIOU35 0.290	AIOU35 0.050	AIO25 0.102	EIO35 0.204	EIOU15 1.168	AIU12 0.778
…						
156.	AEIU23 0.000	AEI12 0.000	AEIU15 0.000	EOU35 0.000	EOU23 0.006	EIOU23 0.012
157.	AOU13 0.000	AEIU14 0.000	AEO12 0.000	AEIOU25 0.000	AEU24 0.005	AEOU14 0.007
158.	AEI12 0.000	AOU25 0.000	AEU25 0.000	AIOU25 0.000	AEIOU24 0.003	AOU25 0.005
159.	AEO13 0.000	AIO15 0.000	AOU25 0.000	EIO25 0.000	AEI12 0.003	AIOU25 0.004
160.	EOU23 0.000	AEIOU13 0.000	IOU14 0.000	AIOU12 0.000	AEO34 0.002	AIOU12 0.001

6.4. Closure of vowel polygons

In this section were presented differences within speakers as well as differences within vowel polygon parameters and their mutual correlation between normal speech and stress influence. The ExamStress database [121] was used in introduced experiments and further divided into two groups- middle and high stress level. These two different groups were finally merged together for creating mixed stress group. The relationships between observed parameter couples were observed by cross-correlation coefficient and statistical parameter called variation coefficient (R) for investigating the suitability of reached result over formant planes and vowel shapes. By these observations was achieved fact that means methods (method 5 and 6) do not reached the highest cross-correlation values but are the most suitable over all vowel shapes and formant planes. This fact is validated for all stress groups and comparing to other method its biggest impact is for mixed stress.

Furthermore, the appropriateness for possibly stress detection was classified by created efficiency coefficient based on classic efficiency equation for each individual vowel polygon separately. By this pointer few statements can be laid. Methods 1, 5 and 6 reached best results in all stress groups, but only method 5 can be used for stress classification. This fact is caused by reached E_c results lower than 1 of other methods. Huge E_c differences within stress groups were also investigated in this paper, exactly the E_c value decreases with increasing stress level within all used methods. Due to this reason, it is convenient to develop stress level pre-classifier or observe unknown stress level influenced speech as middle and high stress level separately for better results achievement.

It was proved some vowel polygons are not suitable for stress detection due to their low cross-correlation value and low uniformity in shape and formant plane criterion. It was also proofed that the lower formant planes contain foremost information about spoken phoneme and higher information of speaker's state and identity are attenuated. As the best vowel shape proves to be AIO, AIU, AEU vowel triangles and AEIU, AEIO and EIOU vowel tetragons for stress detection. Obviously, the best formant planes for stress detection are F1F5, F2F5 and F3F5. In conclusion, stress can be possibly uncovered by usage of mentioned vowel shapes and formant planes (leaded to a various number of vowel polygons) by the fifth experimental method. Obtained results can be further practically applied in call centres, customer services, hospital and security facilities, etc.

It was also uncovered that all vowel polygons are not suitable for speaker recognition, but the most uniform vowel polygons within speaker database can represent the normal state of speaker as the best.

7. Formant changes varying on emotions

Generally, formants are sensitive on actual speaker's state (not only on stress). Due to this fact, this section briefly presents background results achieved in the field of simulated emotion recognition, exactly only on observations in formants positions occurred in influenced speech.

At the beginning, it is necessary to mention the existing research was almost oriented on the speaker recognition described in previous subsection.

Firstly acquired results in the field of emotion recognition were presented at student conference [132]. Used method was based on the changes of important LPC points relative positions between two neighbouring formants. For investigation of changes in frequency domain are observed three frequency bands separately bounded by consecutive formants. Between both bounding formats lie lower formant top point of inflection, antiformant (the minimum) and bottom point of inflection for higher formant. It has to be noticed that the bottom point of inflection IFB1 and the top point of inflection IFT4 are not observed because they lie out of investigated bands (see Fig. 32). The positions of important points are calculated by the first, respectively by the second, derivation of LPC spectrum.

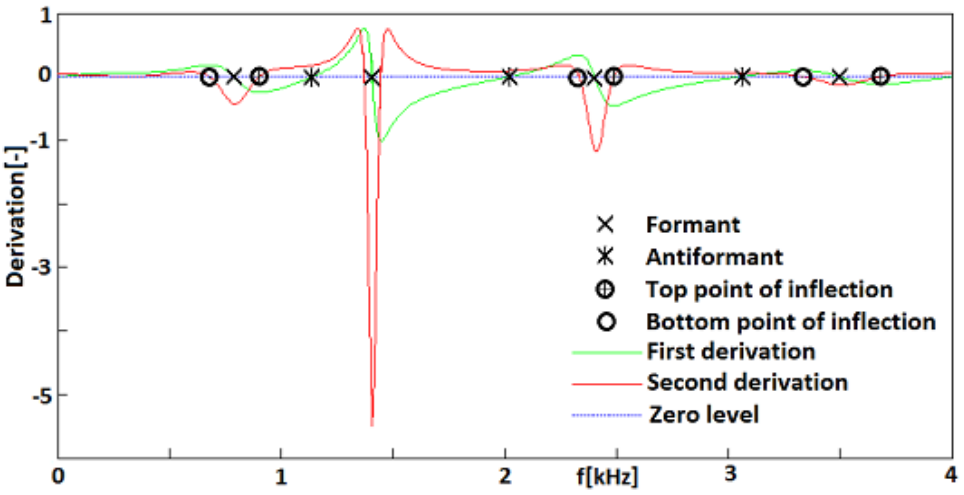

Fig. 32 The derivations of /a/ vowel LPC spectrum and its important points.

Relative position *length* of important point in actual frequency interval is calculated as

$$length_X[\%] = \frac{X - A}{B - A} \cdot 100, \tag{12}$$

where X is frequency of current important point (e.g. the point of inflection, antiformant), A is frequency of lower formant and B is higher formant frequency in observed interval. All calculated relative positions are further displayed by relevant histograms for better observation of the differences between emotional states.

For one speaker, the voluntary theatre actor, were recorded emotional states like normal mood, anger, sadness, happiness and alcohol intoxication. All of these recorded emotional

states were simulated. From relevant records were extracted useful speech features for all vowels. In Fig. 33, results reached for only for /a/ vowel are illustrated.

Results obtained by speech signal analysing are statistically processed for better changes illustration. Each investigated frequency band contains three histograms of extracted results. For better resolution are colours of generated histograms within obtained frequency band different. The lower formant top point of inflection is green, histogram of antiformant is red and the bottom point of inflection of higher formant is blue. In Figure 33, histograms obtained for simulated emotional states in formant bands F1-F2 and F2-F3 are illustrated.

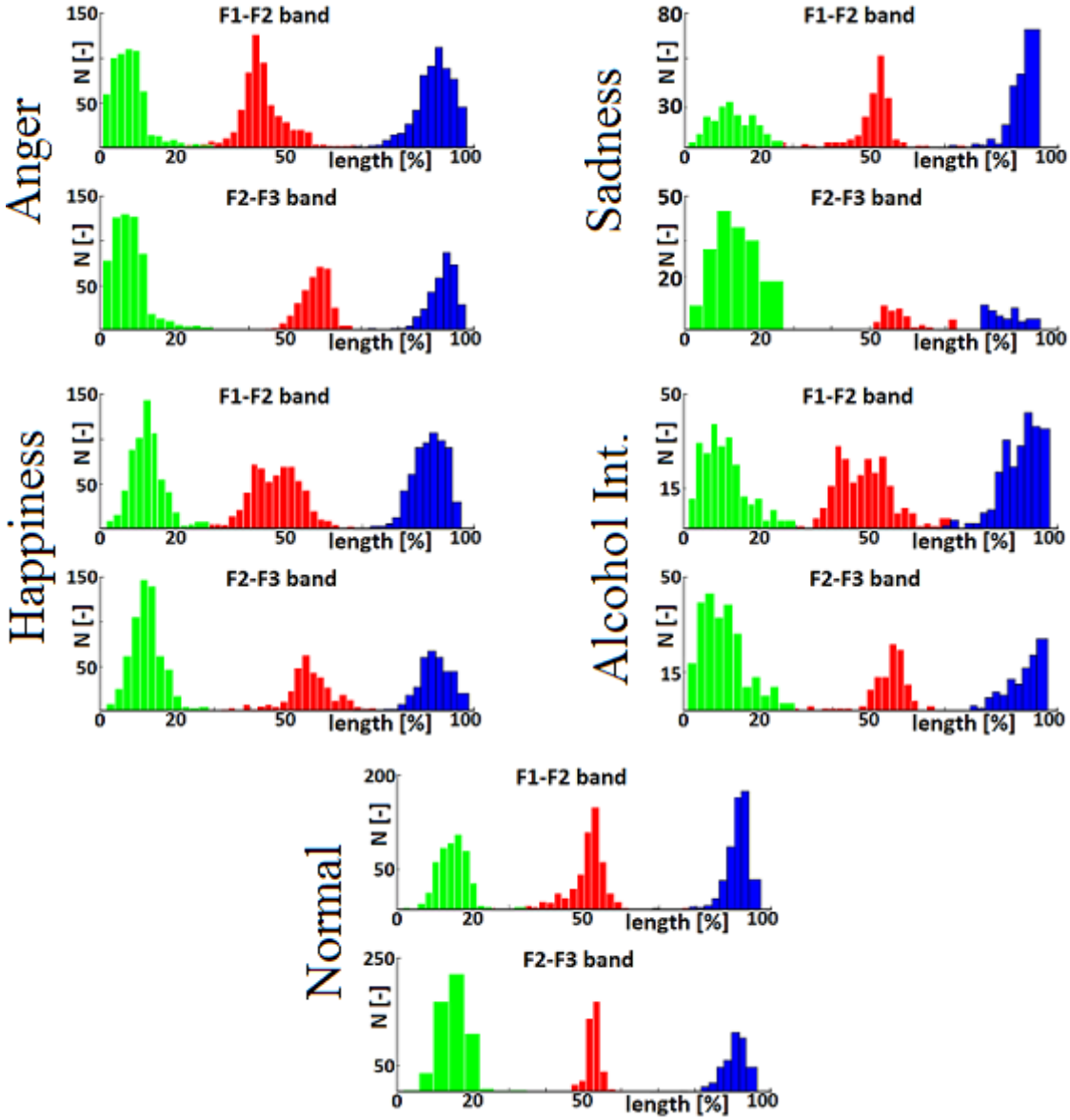

Fig. 33 Relative positions of important points in F1-F2 and F2-F3 bands.

The differences of each individual histogram shape depending on emotional state of speaker are evident. Narrow histograms and high rate of relative positions for each important point are typical for normal state of speaker. Anger can be classified also by high rate of relative positions of important points but its histograms are wider than for normal mood. In the case of anger, the centre of histogram for the first antiformant is shifted from 55% (normal mood) to 40% and the histogram of the second antiformant has absolutely different shape

which is similar to histogram shape of bottom point of inflection of higher formant. Modus of histogram for the second antiformant influenced by anger has relative position 60%.

For sadness, the less rate of each point relative position is typical. The least rate was reached by the second antiformant and the bottom point of inflection. In the other case the states for happiness and alcohol intoxication are more a less similar. Both states have wide histograms for each important point. The shapes of antiformants are related to normal (Gaussian) distribution. The difference between these moods is in the shape of histogram for points of inflection. For happiness, the shape of histogram for both points of inflection has Ricean distribution and for alcohol intoxication is the shape of same histograms similar to Rayleigh distribution. The rate of each relative position is higher for happiness. Achieved results for other vowels are similar to results obtained for /a/ vowel with significant differences of histogram shapes within the emotional state database.

References

[1] S.J.C. GAULIN AND D.H. MCBURNEY, *Evolutionary Psychology*. Boston: Prentice Hall, 2003, ISBN 978–0–13–111529–3.

[2] C. DARWIN, *The Expression of the Emotions in Man and Animals*. London: John Murray, 1872.

[3] E. FOX, *Emotion Science: An Integration of Cognitive and Neuroscientific Approaches to Understanding Human Emotions*. Palgrave MacMillan, 2008, ISBN 978–0–230–00517–4.

[4] P. EKMAN, "An Argument for Basic Emotions," *Cognition & Emotion*, vol. 89, no. 4, pp. 344–350, 2001.

[5] R. PLUTCHIK, "The Nature of Emotions," *American Scientist*, vol. 6, no. 3–4, pp. 169–200, 1992.

[6] D.L. SCHACTER, *Psychology*. New York: Worth Publishers, 2011, ISBN 978–1–4292–3719–2.

[7] S.G. KOOLAUGUDI AND S.R. KROTHAPALLI, "Emotion Recognition from Speech: A Review," *International Journal on Speech Technology*, vol. 15, no. 2, pp. 99–117, 2012.

[8] S.R. KROTHAPALLI AND S.G. KOOLAUGUDI, "Characterization and Recognition of Emotions from Speech Using Excitation Source Information," *International Journal on Speech Technology*, vol. 16, no. 2, pp. 181–201, 2013.

[9] A. BAJPAI AND B. YEGNANARAYANA, "Combining Evidence from Subsegmental and Segmental Features for Audio Clip Classification," in *Proc. IEEE Region 10 Conference (TENCON) 2008*. Hyderabad (India), 2008, pp. 1–5.

[10] D. VERVERIDIS AND C. KOTROPOULOS, "Emotional Speech Recognition: Resources, Features and Methods," *Speech Communication*, vol. 48, pp. 1162–1181, 2006.

[11] M.E. AYADI, M.S. KAMEL, AND F. KARRAY, "Survey on Speech Emotion Recognition: Features, classification Schemes and Databases," *Pattern Recognition*, vol. 44, pp. 572–587, 2011.

[12] T. NWE, S. FOO, AND L. DE SILVA, "Speech Emotion Recognition Using Hidden Markov Models," *Speech Communication*, vol. 41, pp. 603–623, 2003.

[13] H. TEAGER, "Some Observations on Oral Air Flow During Phonation," *IEEE Trans. on Acoust. Speech Signal Process*, vol. 28, no. 5, pp. 599–601, 1990.

[14] J. KAISER, "On a Simple Algorithm to Calculate the 'energy' of a Signal," in *ICASSP-90*, vol. 1, pp. 381–384, 1990.

[15] H. TEAGER AND S. TEAGER, "Evidence for Nonlinear Production Mechanisms in the Vocal Tract," *Speech Production and Speech Modelling, NATO Advanced Institute*, vol. 55, pp. 241–261, 1990.

[16] B. SCHULLER, A. BATLINER, S. STEIDL, AND D. SEPPI, "Recognising Realistic Emotions and Affect in Speech: State of the Art and Lessons Learnt from First Challenge," *Speech Communication*, vol. 53, pp. 1062–1087, 2011.

[17] A. NOGUEIRAS, J.B. MARINO, A. MORENO, AND A. BONAFONTE, "Speech Emotion Recognition Using Hidden Markov Models," in *Proc. European Conf. on Speech Communication and Technology (Eurospeech '01)*. Aalborg (Denmark), 2001, pp. 2679–2682.

[18] B. SCHULLER, G. RIGOLL, AND M. LANG, "Hidden Markov Model-Based Speech Emotion Recognition," in *Proc. ICASSP 2003*. Hong Kong (China), 2003, pp. 1–4.

[19] J.S. DEVI, Y. SRINIVAS, AND S.D. NANDYALA, "Automatic Speech Emotion and Speaker Recognition Based on Hybrid GMM and FFBNN," *International Journal on Computational Sciences & Applications*, vol. 4, no. 1, pp. 35–42, 2014.

[20] M.M.H. EL AYADI, M.S. KAMEL, AND F. KARRAY, "Speech Emotion Recognition Using Gaussian Mixture Vector Autoregressive Models," in *Proc. IEEE International Conf. on Acoustics, Speech and Signal Processing (ICASSP) 2007*. Honolulu (Haiti), 2007, pp. 957–960.

[21] A. PAESCHKE AND W.F. SENDLMEIER, "Prosodic Characteristics of Emotional Speech: Measurements of Fundamental Frequency Movements," in *Proc. ISCA Workshop on Speech and Emotion*. Belfast (United Kingdom), 2000, pp. 75–80.

[22] E.M. ALBORNOZ, D.H. MILONE, AND H.L. RUFINER, "Spoken Emotion Recognition Using Hierarchical Classifiers," *Computer Speech & Language*, vol. 25, no. 3, pp. 556–570, 2011.

[23] S. WU, T.H. FALK, AND W. CHAN, "Automatic Speech Emotion Recognition Using Modulation Spectral Features," *Speech Communication*, vol. 53, no. 5, pp. 768–785, 2011.

[24] P. SHEN, Z. CHANGJUN, AND X. CHEN, "Automatic Speech Emotion Recognition Using Support Vector Machine," in *Proc. International Conf. on Electronic and Mechanical Engineering and Information Technology (EMEIT) 2011*. Harbin (China), 2011, pp. 621–625.

[25] B. SCHULLER, G. RIGOLL, AND M. LANG, "Speech Emotion Recognition Combining Acoustic Features and Linguistic Information in a Hybrid Support Vector Machine-Belief Network Architecture," in *Proc. IEEE International Conf. on Acoustics, Speech and Signal Processing (ICASSP '04) 2004*. 2004, pp. 577–580.

[26] B. VLASENKO, D. PRYLIPKO, R. BÖCK, AND A. WENDEMUTH, "Modeling Phonetic Pattern Variability in Favor of the Creation of Robust Emotion Classifiers for Real-Life Applications," *Computer Speech & Language*, vol. 28, no. 2, pp. 483–500, 2014.

[27] A.P. SIMPSON, K.J. KOHLER, ET AL., *The Kiel Corpus of Read/Spontaneous Speech*: Acoustic database, processing tools and analysis results. 1996.

[28] M. GRIMM, K. KROSCHEL, AND S. NARAYANAN, "The Vera am Mittag German Audio-Visual Emotional Speech Database," in *Proc. IEEE International Conf. on Multimedia and Expo (ICME)*. Hanover (Germany), 2008.

[29] F. BURKHARDT, A. PAESCHKE, ET AL., "A Database of German Emotional Speech," in *Proc. INTERSPEECH 2005*. 2005, pp. 1517–1520.

[30] Z. ZIXING, F. WENIGER, M. WOLLMER, AND B. SCHULLER, "Unsupervised Learning in Cross-Corpus Acoustic Emotion Recognition," in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. Waikoloa (Haiti), 2011, pp. 523–528.

[31] D. JUNG, Z. ZIXING, E. MARCHI, AND B. SCHULLER, "Sparse Autoencoder-Based Feature Transfer Learning for Speech Emotion Recognition," in *Proc. 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII)*. Geneva (Switzerland), 2013, pp. 511–516.

[32] A. STUHLSATZ, C. MEYER, ET AL., "Deep Neural Networks for Acoustic Emotion Recognition: Raising the Benchmarks," in *Proc. IEEE International Conf. on Acoustics, Speech and Signal Processing (ICASSP) 2011*. Prague (Czech Republic), 2011, pp. 5688–5691.

[33] D. GHARAVIAN, M. SHEIKHAN, A. NAZARIEH, AND S. GAROUCY, "Speech Emotion Recognition Using FCBF Feature Selection Method and GA-Optimized Fuzzy ARTMAP Neural Network," *Neural Computing & Applications*, vol. 21, no. 8, pp. 2115–2126, 2012.

[34] E. BOZKURT, E. ERZIN, ET AL., "Formant Position Based Weighted Spectral Features for Emotion Recognition," *Speech Communication*, vol. 53, no. 9–10, pp. 1186–1197, 2011.

[35] L. CHEN, X. MAO, Y. XUE, AND L.L. CHENG, "Speech Emotion Recognition: Features and Classification Models," *Digital Signal Processing*, vol. 22, no. 6, pp. 1154–1160, 2012.

[36] C.C. LEE, E. MOWER, C. BUSSO, S. LEE, AND S. NARAYANAN, "Emotion Recognition Using a Hierarchical Binary Decision Tree Approach," *Speech Communication*, vol. 53, no. 9–10, pp. 1162–1171, 2011.

[37] A. BATLINER, C. HACKER, S. STEIDL, ET AL., "You Stupid Tin Box – Children Interacting with the AIBO Robot: A Cross-Linguistic Emotional Speech Corpus," in *Proc. 4th International Conf. of Language Resources and Evaluation (LREC) 2004*. Lisbon (Portugal), 2004, pp. 171–174.

[38] C. BUSSO, M. BULUT, ET AL., "IEMOCAP: Interactive Emotional Dyadic Motion Capture Database," *Language Resources & Evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[39] V. GARG, H. KUMAR, AND R. SINHA, "Speech Based Emotion Recognition Based on Hierarchical Decision Tree with SVM, BLG and SVR Classifiers," in *Proc. National Conf. on Communications (NCC) 2013*. New Delhi (India), 2013, pp. 1–5.

[40] D. MORRISON, R. WANG, AND L.C. DE SILVA, "Ensemble Methods for Spoken Emotion Recognition in Call-Centres," *Speech Communication*, vol. 49, no. 2, pp. 98–112, 2007.

[41] C. WU AND W. LIANG, "Emotion Recognition of Affective Speech Based on Multiple Classifiers Using Acoustic-Prosodic Information and Semantic Labels," *IEEE Trans. on Affective Computing*, vol. 2, no. 1, pp. 10–21, 2011.

[42] S.G. KOOLAGUDI, N. KUMAR, AND K.S. RAO, "Speech Emotion Recognition Using Segmental Level Prosodic Analysis," in *Proc. International Conf. on Devices and Communications (ICDeCom) 2011*. 2011, pp. 1–5.

[43] K.S. RAO, S.G. KOOLAUGUDI, AND R.R. VEMPADA, "Emotion Recognition from Speech Using Global and Local Prosodic Features," *International Journal on Speech Technology*, vol. 16, no. 2, pp. 143–160, 2013.

[44] D. POVEY, A. GHOSHAL, ET AL., "The Kaldi speech Recognition Toolkit," in *Proc. IEEE ASRU*. 2011.

[45] S.G. KOOLAUGUDI, AND R.S. KROTHAPALLI, "Two Stage Emotion Recognition Based on Speaking Rate," *International Journal on Speech Technology*, vol. 14, no. 1, pp. 35–48, 2011.

[46] S. JOHAR, "Paralinguistic Profiling Using Speech Recognition," *International Journal on Speech Technology*, vol. 17, no. 3, pp. 205–209, 2014.

[47] Q. MAO, X. ZHAO, Z. HUANG, AND Y. ZHAN, "Speaker-Independent Speech Emotion Recognition by Fusion of Functional and Accompanying Paralanguage Features," *Journal of Zhejiang University SCIENCE C*, vol. 14, no. 7, pp. 573–582, 2013.

[48] W. YE AND X. FAN, "Bimodal Emotion Recognition from Speech and Text," International *Journal of Advanced Computer Science and Applications*, vol. 5, no. 2, 2014.

[49] S. NTALAMPIRAS AND N. FAKOTAKIS, "Modeling the Temporal Evolution of Acoustic Parameters for Speech Emotion Recognition," *IEEE Trans. on Affective Computing*, vol. 3, no. 1, pp. 116–125, 2012.

[50] J. DENG, W. HAN, AND B. SCHULLER, "Confidence Measures for Speech Emotion Recognition: A Start," in *Proc. 10th ITG Symposium on Speech Communication*. Braunschweig (Germany), 2012, pp. 1–4.

[51] S.G. KOOLAUGUDI, AND R.S. KROTHAPALLI, "Emotion Recognition from Speech Using Sub-Syllabic and Pitch Synchronous Spectral Features," *International Journal on Speech Technology*, vol. 15, no. 4, pp. 495–511, 2012.

[52] C. BUSSO, A. METALLINOU, AND S.S. NARAYANAN, "Iterative Feature Normalization for Emotional Speech Detection," in *Proc. IEEE International Conf. on Acoustics, Speech and Signal Processing (ICASSP) 2011*. Prague (Czech Republic), 2011, pp. 5692–5695.

[53] R. VAN BEZOOIJEN, *The Characterictics and Recognizability of Vocal Expression of Emotions*, 1984.

[54] P. ROACH, "Techniques for the Phonetic Description of Emotional Speech," in *Proc. ISCA Workshop on Speech and Emotion*. Belfast (United Kingdom), 2000, pp. 53–59.

[55] S. MCGILLOWAY, *Negative Symptoms and Speech Parameters in Schizophrenia.* Unpublished doctoral thesis. Queen's University Belfast (United Kingdom), 1997.

[56] I.S. ENGBERG ET AL., "Design, Recording and Verification of a Danish Emotional Speech Database," in *Proc. EUROSPEECH '97*. 1997, pp. 1695–1698.

[57] S. MOZZICONACCI, *Speech Variability and Emotion: Production and Perception.* Unpublished doctoral thesis. Technical University Eindhoven (Netherlands), 1998.

[58] A. ABELIN AND J. ALLWOOD, "Cross Linguistic Interpretation of Emotional Prosody," in *Proc. ISCA Workshop on Speech and Emotion*. Newcastle (United Kingdom), 2000, pp. 110–113.

[59] E. DOUGLAS-COWIE, R. COWIE, AND M. SCHROEDER, "HUMAINE," in *Proc. ISCA Workshop on Speech and Emotion*. Belfast (United Kingdom), 2000.

[60] D. FRANCE ET AL., "Acoustical Properties of Speech as Indicators of Depression and Suicidal Risk," *IEEE Trans. on Biomed. Eng.*, vol. 47, no. 7, 2000.

[61] I. IRIONDO ET AL., "Validation of an Acoustical Modeling of Emotional Expression in Spanish Using Speech Synthesis Techniques," in *Proc. ISCA Workshop on Speech and Emotion*. Belfast (United Kingdom), 2000, pp. 161–166.

[62] C. PEREIRA, "Dimensions of Emotional Meaning in Speech," in *Proc. ISCA Workshop on Speech and Emotion*. Newcastle (Northern Ireland), 2000, pp. 25–28.

[63] N. CAMPBELL, "Recording and Storing of Speech Data," in *Proc. LREC 2002*. Las Palmas (Canary Islands), 2002.

[64] J. ANG ET AL., "Prosody-Based Automatic Detection of Annoyance and Frustration in Human-Computer Dialog," in *Proc. ICSLP 2002*. Denver (Colorado, USA), 2002.

[65] A. BATLINER ET AL., "How to Finf Trouble in Communication," *Speech Communication*, vol. 40, pp. 117–143, 2003.

[66] S. YACOUB ET AL., "Recognition of Emotions in Interactive Voice Response Systems," in *Proc. EUROSPEECH 2003*. Geneva (Switzerland), 2003, pp. 1–4.

[67] S.T. JOVICIC ET AL., "Corpus Creating of Speech Expression of Emotions and Attitudes in Serbian Language - GEES," in *Proc. Conference TELFOR 2003*. Belgrade (Serbia), 2003.

[68] EUROPEAN LANGUAGE RESOURCES ASSOCIATION, *Groningen ELRA Corpus Number S0020*, 2005.

[69] G. BELLER ET AL., "Hybrid Concatenative Synthesis in the Intersection of Speech and Music," in *Proc. JIM2005*. Paris (France), 2005, pp. 41–45.

[70] S. YILDIRIM, S. NARAYANAN, AND A. POTAMIANOS, "Detecting Emotional State of Child in a Conversational Computer Game," *Computer Speech & Language*, vol. 25, no. 1, pp. 29–44, 2011.

[71] V. PETRUSHIN, "Emotion Recognition in Speech Signal: Experimental Study, Development, and Application," in *Proc. ICSLP 2000*. Denver (USA), 2000.

[72] J. WAGNER, F. LINGENFELSER, AND E. ANDRE, "The Social Signal Interpretation Framework (SSI) for Real Time Signal Processing and Recognition," in *Proc. 12th Annual Conference of the International Speech Communication Association INTERSPEECH 2011*. Florence (Italy), 2011, pp. 3245–3248.

[73] T. VOGT, E. ANDRE, AND N. BEE, "EmoVoice – A Framework for Online Recognition of Emotions from Voice," *Perception in Multimodal Dialogue Systems*, pp. 188–199, 2008.

[74] EMOSPEECH, EmoSpeech [Computer Program], 2014, http://www.emospeech.net/.

[75] P. BOERSMA AND D. WEENINK, Praat: Doing Phonetics by Computer [Computer Program], 2014, http://www.praat.org/.

[76] N. SHARMA AND T. GEDEON, "Objective Measures, Sensors and Computational Techniques for Stress Recognition and Classification: A Survey," *Computer Methods and Programs in Biomedicine*, vol. 108, no. 3, pp. 1287–1301, 2012.

[77] M. LECH AND L. HE, "Stress and Emotion Recognition Using Acoustic Speech Analysis," *Mental Health Informatics*, pp. 163–184, 2014.

[78] B. VLASENKO, B. SCHULLER ET AL., "Balancing Spoken Content Adaptation and Unit Length in the Recognition of Emotion and Interest," in *Proc. Conference of the International Speech Communication Association INTERSPEECH 2008*. Brisbane (Australia), 2008, pp. 805–808.

[79] J.H.L. HANSEN AND S. PATIL, "Speech under Stress: Analysis, Modeling and Recognition," *Speaker Classification 1*, pp. 108–137, 2007.

[80] L. HE, M. LECH, N.C. MADDAGE, AND N.B. ALLEN, "Study of Empirical Mode Decomposition and Spectral Analysis for Stress and Emotion Classification in Natural Speech," *Biomedical Signal Processing and Control*, vol. 6, no. 2, pp. 139–146, 2011.

[81] M.H. FAROUK, "Emotion Recognition from Speech," *Application of Wavelets in Speech Processing*, pp. 31–32, 2014.

[82] B. YANG AND M. LUGGER, "Emotion Recognition from Speech Signals Using New Harmony Features," *Signal Processing*, vol. 90, no. 5, pp. 1415–1423, 2010.

[83] M. SIGMUND, "Spectral Analysis of Speech under Stress," *International Journal of Computer Science and Network Security*, vol. 7, no. 4, pp. 170–172, 2007.

[84] M. SIGMUND, "Statistical Analysis of Fundamental Frequency Based Features in Speech under Stress," *Information and Technology Control*, vol. 42, no. 3, pp. 286–291, 2013.

[85] L.A. STREETER ET AL., "Pitch Changes during Attempted Deception," *Journal on Personality and Social Psychology*, vol. 35, no. 5, pp. 345–350, 1977.

[86] S. MOON AND B. LINDBLOM, "Interaction between Duration, Context, and Speaking Style in English Stressed Vowels," *Journal of the Acoustical Society of America*, vol. 96, no. 40, 1994.

[87] G. ZHOU, J.H.L. HANSEN, ET AL., "Nonlinear Feature Based Classification of Speech under Stress," *IEEE Trans. on Speech and Audio Processing*, vol. 9, no. 3, pp. 201–216, 2001.

[88] A.I. ILIEV, M.S. SCORDILIS, J.P. PAPA, AND A.X. FALCO, "Spoken Emotion Recognition through Optimum-Path Forest Classification Using Glottal Features," *Computer Speech & Language*, vol. 24, no. 3, pp. 445–460, 2010.

[89] THE NATO, SUSC: Speech Under Stress Databases [Speech Database], http://cslu.colorado.edu/rspl/stress.html/.

[90] F.J. TOLKMITT AND K.R. SCHERER, "Effect of Experimentally Induced Stress on Vocal Parameters," *Journal of Experimental Psychology*, vol. 12, no. 3, pp. 302–313, 1986.

[91] J.H.L. HANSEN AND S. BOU-GHAZALE, "Getting Started with SUSAS: A Speech Under Simulated and Actual Stress Database," in *Proc. EUROSPEECH 1997*. Rhodes (Greece), 1997, pp. 1743–1746.

[92] M. RAHURKAR, J.H.L. HANSEN, ET AL., "Frequency Band Analysis for Stress Detection Using Teager Energy Operator Based Feature," in *Proc. ISCA INTERSPEECH '02*. Denver (Colorado, USA), 2002, pp. 2021–2024.

[93] K. SCHERER, "A Cross-Cultural Investigation of Emotion Inferences from Voice and Speech: Implications for Speech Technology," in *Proc. ICSLP 2000*. Beijing (China), 2000.

[94] E. MCMAHON ET AL., "What Chance that a DC Could Recognise Hazardous Mental States from Sensor Outputs?," in *DC Tales*. Santorini (Greece), 2003.

[95] R. FERNANDEZ AND R.W. PICARD, "Modeling Drivers' Speech under Stress," *Speech Communication*, vol. 40, pp. 145–149, 2003.

[96] K. HOFBAUER, S. PETRIK, AND H. HERING, "The ATCOSIM Corpus of Non-Prompted Clean Air Traffic Control Speech," in *Proc. LREC 2008*. Marrakech (Morocco), 2008, pp. 2147–2152.

[97] J. LUIG, Report 45/11: The IEM Pilot Speech Database [Technical Report], Institute of Electronic Music and Acoustics, 2011.

[98] V.P. PATIL, K.K. NAYAK, AND M. SAXENA, "Voice Stress Detection," International *Journal of Electrical, Electronics and Computer Engineering*, vol. 2, no. 2, pp. 148–154, 2013.

[99] P. PETTA, C. PELACHAUD, AND R. COWIE, *Emotion-Oriented Systems: The Humaine Handbook*. Berlin: Springer, 2011, ISBN 978–3–642–15183–5.

[100] NEMESYSCO, LVA, SENSE & LioNet Technology [Computer Program], 2014, http://www.nemesysco.com/.

[101] AGI SENSIBILITY TECHNOLOGY, X13-VSA [Computer Program], 2014, http://www.lie-web.co.jp/.

[102] X13-VSA, The X13-VSA Home/PRO/Cobra [Computer Program], 2014, http://www.lie-detection.com/.

[103] AVSAPRO, The AVSAPRO: Voice Stress Analysis Systems [Computer Program], 2014, http://www.avsapro.com/.

[104] S. TOKUNO, G. TSUMATORI, S. SHONO, ET AL., "Usage of Emotion Recognition in Military Health Care," in *Proc. Defense Science Research Conference and Expo (DSR) 2011*. Singapore (Singapore), 2011, pp. 1–55.

[105] M. LEWIS, J.M. HAVILAND-JONES, AND L.F. BARRETT, *Handbook of Emotions*. New York: The Guilford Press, 2008, ISBN 978–1–59385–650–2.

[106] H. SCHLOSBERG, "Three Dimensions of Emotion," *Psychological Review*, vol. 61, no. 2, pp. 81–88, 1954.

[107] R. PLUTCHIK, *Emotion: A Psychoevolutionary Synthesis*. Harper & Row, 1980, ISBN 978–0–06045–235–3.

[108] S.E. AMBROSE AND R.W. RAND, The Bruce McPherson Infrasound and Low Frequency Noise Study: Adverse Healt Effects Produced by Large Industrial Wind Turbines Confirmed [Study Report], INCE, 2011.

[109] GRAS, G.R.A.S. 40AN: ½" Free-Field Mic [Infrasonic Microphone], 2014, http://www.gras.dk/40an.html.

[110] M. WRZOSEK, J. MACULEWICZ, ET AL., "Pitch Processing of Speech: Comparison of Psychoacoustic and Electrophysiological Data," *Archives of Acoustics*, vol. 38, no. 3, pp. 375–381, 2013.

[111] J. PSUTKA, L. MÜLLER, J. MATOUŠEK, AND V. RADOVÁ, *Mluvíme s počítačem česky*. Prague: Academia, 2006, ISBN 80-200-1309-1, in Czech.

[112] M. STANĚK, "Software System for Speech Signal Processing," in *Proc. 17th International Student Conference on Electrical Engineering POSTER 2013*. Prague (Czech Republic), 2013, pp. 1-5.

[113] M. STANĚK AND L. POLÁK, "Algorithms for Vowel Recognition in Fluent Speech Based on Formant Positions," in *Proc. 36th International Conference on Telecommunication and Signal Processing (TSP) 2013*. Rome (Italy), 2013, pp. 521–525.

[114] R.C. SCHNELL AND F. MILINAZZO, "Formant Location from LPC Analysis Data," *IEEE Transactions on Speech and Audio Processing*, vol. 1, pp. 129–134, 1993.

[115] M. STANĚK, "Software for Generation and Analysis of Vowel Polygons," in *Proc. 37th International Conference on Telecommunication and Signal Processing (TSP) 2014*. Berlin (Germany), 2014, pp. 424–427.

[116] M. STANĚK AND M. SIGMUND, "Porovnání efektivity řečových spektrálních parametrů pro identifikaci mluvčích," *Elektrorevue- Internetový časopis,* vol. 2013, no. 8, pp. 1–8, 2013, in Czech.

[117] M. STANĚK AND M. SIGMUND, "Speaker Dependent Changes in Formants Based on Normalization of Vowel Triangle," in *Proc. 23th International Conference RADIOELEKTRONIKA 2013*. Pardubice (Czech Republic), 2013, pp. 337–341.

[118] M. STANĚK AND M. SIGMUND, "Comparison of Speaker Individuality in Triangle Areas of Plane Formant Spaces," in *Proc. 24th International Conference RADIOELEKTRONIKA 2014*. Bratislava (Slovakia), 2014, pp. 1–4.

[119] M. SIGMUND AND R. MENŠÍK, "Estimation of Vocal Tract Long-Time Spectrum," *ITG-Fachberichte*, no. 152, pp. 69–71, 1998.

[120] M. STANĚK AND M. SIGMUND, "Speaker Distinction Using Vowel Polygons: Experimental Study," in *Proc. 25th International Conference RADIOELEKTRONIKA 2015*. Pardubice (Czech Republic), 2015, pp. 125–128.

[121] M. SIGMUND, "Introducing the Database ExamStress for Speech Under Stress," in *Proc. 7th Nordic Signal Processing Symposium NORSIG 2006*. Reykjavik (Iceland), 2006, pp. 290–293.

[122] M. STANĚK AND M. SIGMUND, "Finding the Most Uniform Vowel Polygon Behavior Caused by Psychological Stress Influence," *Radioengineering*, vol. 24, no. 2, pp. 604–609, 2015.

[123] M. STANĚK, "Vowel Polygon Efficiency Induced by Middle Level Psychological Stress," in *Proc. 38th International Conference on Telecommunication and Signal Processing (TSP) 2015*. Prague (Czech Republic), 2015, pp. 387–391.

[124] M. SIGMUND, "Influence of psychological stress on formant structure of vowels," *Electronics and Electrical Engineering*, vol. 18, no. 10, pp. 45 – 48, 2012.

[125] S. CHANDAKA, A. CHATTERJEE, AND S. MUNSHI, "Support vector machines employing cross-correlation for emotional speech recognition," *Measurement*, vol. 42, no. 4, pp. 611 – 618, 2009.

[126] A. G. ADAMI AND H. HERMANSKY, "Segmentation of speech for speaker and language recognition," in *Proc. EUROSPEECH 2003*. Geneva (Switzerland), 2003, pp. 1 – 4.

[127] S. N. WRIGLEY, G. J. BROWN, V. WAN, AND S. RENALS, "Speech and crosstalk detection in multichannel audio," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 1, pp. 84-91, 2005.

[128] S. ROSEN AND E. MANGANARI, "Is there a relationship between speech and nonspeech auditory processing in children with dyslexia?," *Journal of Speech, Language and Hearing Research*, vol. 44, pp. 720 – 736, 2001.

[129] S. YILDIRIM, M. BULUT, C. M. LEE, A. KAZEMZADEH, C. BUSSO, Z. DENG, S. LEE, AND S. NARAYANAN, "An acoustic study of emotions expressed in speech," in *Proc. International Conference on Spoken Language Processing*. Jeju Island (Korea), 2004, pp. 2193 - 2196.

[130] S. Y. OH AND K. Y. CHUNG, "Target speech feature extraction using non-parametric correlation coefficient," *Cluster Computing*, vol. 17, no. 3, pp. 893 – 899, 2014.

[131] M. DIETRICH AND K. V. ABBOTT, "Psychobiological stress reactivity and personality in persons with high and low stressor-induced extralaryngeal reactivity (to be published)," *Journal of Speech, Language and Hearing Research*, 2014.

[132] M. STANĚK, "Method for Recognition the Actual Emotional State of Speaker," in *Proc. 19th Student Conference EEICT 2013*. Brno (Czech Republic), 2013, pp. 55–59.

[133] J. P. CABRAL, K. RICHMOND, J. YAMAGISHI, AND S. RENALS, "Glottal Spectral Separation for Speech Synthesis," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 2, pp. 195-208, 2014.

[134] T. RAITIO, A. SUNI, J. YAMAGISHI, H. PULAKKA, J. NURMINEN, M. VAINIO, AND P. ALKU, "HMM-Based speech synthesis utilizing glottal inverse filtering," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 19, no. 1, pp. 153–165, 2011.

[135] J. P. CABRAL, S. RENALS, J. YAMAGISHI, AND K. RICHMOND, "HMM Based speech synthesiser using the LF model glottal source," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)2011*, Prague (Czech Republic), 2011, pp. 4704–4707.

[136] F. VILLAVICENCIO, "A strategy for LF based glottal source & vocal tract estimation on stationary modal singing," in *Proc. 22nd European Signal Processing Conference (EUSIPCO)*, Lisbon (Portugal), 2014, pp. 1457–1461.

[137] J. LAVER, *The Phonetic Description of Voice Quality*. Cambridge University Press, UK: Cambridge, 1980.

[138] R. SUN, E. MOORE, AND J. TORRES, "Investigating glottal parameters for differentiating emotional categories with similar prosodics," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Taipei (China), 2009, pp. 4509–4512.

[139] G. SHASHIDHAR, K. KOOLAGUDI, AND K. S. RAO, "Emotion recognition from speech using source, system, and prosodic features," *IEEE Int. Journal of Speech Technology*, vol. 15, no. 2, pp. 265–289, 2011.

[140] A. I. ILIEV AND M. S. SCORDILIS, "Spoken Emotion Recognition Using Glottal Symmetry," *EURASIP Journal on Advances in Signal Processing*, vol. 2011, 2011.

[141] P. K. MONGIA AND R. K. SHARMA, "Estimation and Statistical Analysis of Human Voice Parameters to Investigate the Influence of Psychological Stress and to Determine the Vocal Tract Transfer Function of an Individual," *Journal of Computer Networks and Communications*, vol. 2014, 2014.

[142] I. BISIO, A. DELFINO, F. LAVAGETTO, M. MARCHESE, AND A. SCIARRONE, "Gender-Driven Emotion Recognition Through Speech Signals For Ambient Intelligence Applications", *IEEE Trans. Emerging Topics in Computing*, vol. 1, no. 2, pp. 244–257, 2014.

[143] J. KAMETANI, "*Speaker recognition with glottal pulse-shapes*", U.S. Patent 5091948A, February 25, 1992.

[144] B. YEGNANARAYANA, K. SHARAT REDDY, AND S. P. KISHORE, "Source and system features for speaker recognition using AANN models," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Salt Lake City (USA), 2001, pp. 409–412.

[145] J. WANG AND M. T. JOHNSON, "Psychologically-motivated feature extraction for speaker identification," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Florence (Italy), 2014, pp. 1690–1694.

[146] V. N. SOROKIN, A. A. TANANYKIN, AND V. G. TRUNOV, "Speaker recognition using vocal source model," *Pattern Recognition and Image Analysis*, vol. 24, no. 1, pp. 156–173, 2014.

[147] A. TSANAS, M. A. LITTLE, P. E. MCSHARRY, J. SPIELMAN, AND L. O. RAMING, "Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease," *IEEE Trans. Biomedical Engineering*, vol. 59, no. 5, pp. 1264–1271, 2012.

[148] H. HERZEL, D. BERRY, I. R. TITZE, AND M. SALEH, "Analysis of vocal disorders with methods from nonlinear dynamics," *Journal of Speech, Language, and Hearing Research*, vol. 37, 1994.

[149] M. SIGMUND AND P. ZELINKA, "Analysis of voiced speech excitation due to alcohol intoxication," *Information Technology and Control*, vol. 40, no. 2, pp. 145–150, 2011.

[150] P. GOMEZ-VILDA, O. PEREZ-BROCANO, R. MARTINEZ-OLALLA, V. RODELLAR-BIARGE, K. LOPEZ DE IPINA PENA, M. ECAY, AND P. MARTINEZ-LAGE, "Biomechanical characterization of phonation in Alzheimer's disease," in *Proc. Int. Work Conf. Bio-inspired Intelligence (IWOBI)*, Liberia (Costa Rica), 2014, pp. 14–20.

[151] SALONI, R. K. SHARMA, AND A. K. GUPTA, "Disease detection using voice analysis: a review," *International Journal of Medical Engineering and Informatics*, vol. 6, no. 3, pp. 189–209, 2014.

[152] T. DRUGMAN, P. ALKU, A. ALWAN, AND B. YEGNANARAYENA, "Glottal source processing: From analysis to applications," *Computer Speech & Language*, vol. 28, no. 5, pp. 1117–1138, 2014.

[153] M. AIRAS, "TKK Aparat: an environment for voice inverse filtering and parameterization," *Logopedics, phoniatrics, vocology*, vol. 33, no. 1, pp. 49–68, 2008.

[154] A. LÖFQVIST, "Inverse filtering as a tool in voice research and therapy," *Logopedics, phoniatrics, vocology*, vol. 16, no. 1-2, pp. 8–16, 1991.

[155] M. A. NWACHUKU, "Inverse filtering techniques in speech analysis," *Nigerian Journal of Technology*, vol. 1, no. 1, pp. 38–42, 1991.

[156] S. DIAS AND A. FERREIRA, "Glottal pulse estimation – a frequency domain approach," unpublished.

[157] H. WAKITA, "Direct estimation of the vocal tract shape by inverse filtering of acoustic speech," *IEEE Transactions on Audio and Electroacoustics*, vol. 21, no. 5, pp. 417–427, 1973.

[158] P. ALKU, "Glottal wave analysis with Pitch Synchronous Iterative Adaptive Inverse Filtering," *Speech Communication*, vol. 11, no. 2 3, pp. 109–118, 1992.

[159] G. FANT, J. LILJENCRANTS, AND Q. LIN, "A four-paramter model of glottal flow," *STL-QPSR*, vol. 26, no. 4, pp. 1–13, 1985.

[160] C. ROADS, *Microsound.* Cambridge: The MIT Press, 2001, ISBN 0-262-18215-7.

[161] M. STANĚK AND T. SMATANA, "Comparison of Fundamental Frequency Detection Methods and Introducing Self-Repairing Alhorithm for Musical Applications," in *Proc. 25th International Conference RADIOELEKTRONIKA 2015*. Pardubice (Czech Republic), 2015, pp. 217–221.

[162] M. STANĚK AND M. SIGMUND, "Psychological Stress Detection in Speech Using Return-To-Opening Phase Ratios in Glottis," *Elektronika ir Elektrotechnika*, vol. 21, no. 5, pp. 59–63, 2015.

[163] M. SIGMUND, A. PROKEŠ, AND Z. BRABEC, "Statistical analysis of glottal pulses in speech under psychological stress," in *Proceedings of EUSIPCO*, Lausanne (Switzerland), 2008, pp. 1 5.

[164] S. E. BOU-GHAZALE, "A comparative study of traditional and newly proposed features for recognition of speech under stress," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 4, pp. 429–442, 2000.

[165] D. GHARAVIAN, M. SHEIKHAN, AND F. AHOFTEDEL, "Emotion recognition improvement using normalized formant supplementary features by hybrid of DTW-MLP-GMM model," *Neural Computing & Applications*, vol. 22, no. 6, pp. 1181–1191, 2013.

[166] H. LU, D. FRAUENDORFER, M. RABBI, M. SCHMID MAST, G. T. CHITTARANJAN, A. T. CAMPBELL, D. GATICA-PEREZ, AND T. CHOUDHURY, "StressSense: detecting stress in constrained acoustic environments using smartphones," in *Proceedings of UbiComp '12 –ACM Conference on Ubiquitous Computing*, New York (USA), 2012, pp. 351–360.

[167] M. JESSEN, *Einfluss von Stress auf Sprache und Stimme*. Schulz Kirchner: Idstein, 2006. (in German)

[168] C. KIRCHHUBEL, D. M. HOWARD, AND A. W. STEDMON, "Acoustic correlates of speech when under stress: Research, methods and future directions," *International Journal of Speech, Language and the Law*, vol. 18, no. 1, pp. 75–98, 2011.

[169] A. I. ILIEV, M. S. SCORDILIS, J. P. PAPA, AND A. X. FALCAO, "Spoken emotion recognition through optimum-path forest classification using glottal features," *Computer Speech & Language*, vol. 24, no. 3, pp. 445–460, 2010.

[170] H. MUTHUSAMY, K. POLAT, AND S. YAACOB, "Improved emotion recognition using Gaussian Mixture Model and extreme learning machine in speech and glottal signals," *Mathematical Problems in Engineering*, 2015, pp. 1–13.

[171] D. FISHER, K. CHANG, AND J. CANNY, "A speech analysis library for analyzing affect, stress, and mental health on mobile phones," in *Proceedings of PhoneSense 2011*, Seattle (USA), 2011, pp. 1–5.

[172] M. STANĚK AND M. SIGMUND, "Analysis of Closing-To-Opening Phase Ratio in Top-To-Bottom Glottal Pulse Segmentation for Psychological Stress Detection," *Elektronika ir Elektrotechnika*. Accepted for publication, 2016.

Curriculum Vitae

Name: Miroslav STANĚK

Born: September 13th 1987 in Litoměřice

Contact: mirek.stanek@seznam.cz

Education

2012 – 2016 **Brno University of Technology / Department of Radio Electronics**
Ph.D. study of Electronics and Communication
Dissertation on Stress Recognition from Speech Signal

2010 – 2012 **Brno University of Technology / Department of Radio Electronics**
Master's study of Electronics and Communication
Diploma thesis on Multimedia Signal Processing

2011 – 2013 **Brno University of Technology / Department of Forensic Engineering**
Master's study of Real Estate Engineering

2007 – 2010 **Brno University of Technology / Department of Radio Electronics**
Bachelor's study of Electronics and Communication
Bachelor's thesis on Voltage Source Controlled and Fed by USB Bus

1999 – 2007 **Gymnázium Josefa Jungmanna, Litoměřice**

Experience

8/08 – 12/15 **ALS Czech Republic**
Employee of logistic affairs

4/2016 – **Honeywell Technology Solutions**
Software Design Engineer at Honeywell Aerospace

Languages

English, German

Selected Publications Related to the Doctoral Thesis

Publications in impact journals

M. STANĚK AND M. SIGMUND, Analysis of Closing-To-Opening Phase Ratio in Top-To Bottom Glottal Pulse Segmentation for Psychological Stress Detection. *Elektronika ir Elektrotechnika.* Accepted for publication, 2016. (IF 0.561)

M. STANĚK AND M. SIGMUND, Finding the Most Uniform Vowel Polygon Behavior Caused by Psychological Stress Influence. *Radioengineering*, vol. 24, no. 2, pp. 604–609, 2015. (IF 0.653)

M. STANĚK AND M. SIGMUND, Psychological Stress Detection in Speech Using Return-To-Opening Phase Ratios in Glottis. *Elektronika ir Elektrotechnika*, vol. 21, no. 5, pp. 59–63, 2015. (IF 0.561)

Publications indexed in SCOPUS

M. STANĚK, Vowel Polygon Efficiency Induced by Middle Level Psychological Stress. In *Proceedings of 38th International Conference on Telecommunication and Signal Processing (TSP) 2015*. Prague (Czech Republic), 2015, pp. 387–391.

M. STANĚK AND M. SIGMUND, Speaker Distinction Using Vowel Polygons: Experimental Study. In *Proceedings of 25th International Conference RADIOELEKTRONIKA 2015*. Pardubice (Czech Republic), 2015, pp. 125–128.

M. STANĚK AND M. SIGMUND, Comparison of Speaker Individuality in Triangle Areas of Plane Formant Spaces. In *Proceedings of 24th International Conference RADIOELEKTRONIKA 2014*. Bratislava (Slovakia), 2014, pp. 1–4.

M. STANĚK, Software for Generation and Analysis of Vowel Polygons. In *Proceedings of 37th International Conference on Telecommunication and Signal Processing (TSP) 2014*. Berlin (Germany), 2014, pp. 424-427.

M. STANĚK AND L. POLÁK, Algorithms for Vowel Recognition in Fluent Speech Based on Formant Positions. In *Proceedings of 36th International Conference on Telecommunication and Signal Processing (TSP) 2013*. Rome (Italy), 2013, pp. 521–525.

M. STANĚK AND M. SIGMUND, Speaker Dependent Changes in Formants Based on Normalization of Vowel Triangle. In *Proceedings of 23th International Conference RADIOELEKTRONIKA 2013*. Pardubice (Czech Republic), 2013, pp. 337–341.

Other Publications

M. STANĚK AND T. SMATANA, Comparison of Fundamental Frequency Detection Methods and Introducing Self-Repairing Alhorithm for Musical Applications. In *Proceedings of 25th International Conference RADIOELEKTRONIKA 2015*. Pardubice (Czech Republic), 2015, pp. 217–221.

M. STANĚK AND M. SIGMUND, Porovnání efektivity řečových spektrálních parametrů pro identifikaci mluvčích. *Elektrorevue- Internetový časopis,* vol. 2013, no. 8, pp. 1–8, 2013, in Czech.

M. STANĚK, Software System for Speech Signal Processing. In *Proceedings of 17th International Student Conference on Electrical Engineering POSTER 2013*. Prague (Czech Republic), 2013, pp. 1-5.

M. STANĚK, Method for Recognition the Actual Emotional State of Speaker. In *Proceedings of 19th Student Conference EEICT 2013*. Brno (Czech Republic), 2013, pp. 55–59.

M. STANĚK, Regulovatelný zdroj napájený a řízený pomocí sběrnice USB. In *Proceedings of 16th Student Conference EEICT 2010*. Brno (Czech Republic), 2010, pp. 36–38, in Czech.